Saarland University
Faculty of Philosophy
Department of Philosophy

Building Bridges for Better Machines: From Machine Ethics to Machine Explainability and Back

Doctoral Thesis

for the attainment of the academic degree of

Doctor of Philosophy

of the Faculty of Philosophy of Saarland University

Author

Timo Speith

Supervisor

Prof. Dr. Ulrich Nortmann

Reviewers

Prof. Dr. Ulrich Nortmann

Prof. Dr. Marija Slavkovik

Saarbrücken, 2023

Dean of the Faculty: Stefanie Haberzettl

Day of the Last Examination: June 01, 2023

Still round the corner there may wait
A new road or a secret gate;
And though I oft have passed them by,
A day will come at last when I
Shall take the hidden paths that run
West of the Moon, East of the Sun.

J.R.R. Tolkien, *The Return of the King*

Abstract

Be it nursing robots in Japan, self-driving buses in Germany or automated hiring systems in the USA, complex artificial computing systems have become an indispensable part of our everyday lives. Two major challenges arise from this development: *machine ethics* and *machine explainability*. Machine ethics deals with behavioral constraints on systems to ensure restricted, morally acceptable behavior; machine explainability affords the means to satisfactorily explain the actions and decisions of systems so that human users can understand these systems and, thus, be assured of their socially beneficial effects.

Machine ethics and explainability prove to be particularly efficient only in symbiosis. In this context, this thesis will demonstrate how machine ethics requires machine explainability and how machine explainability includes machine ethics. We develop these two facets using examples from the scenarios above. Based on these examples, we argue for a specific view of machine ethics and suggest how it can be formalized in a theoretical framework.

In terms of machine explainability, we will outline how our proposed framework, by using an argumentation-based approach for decision making, can provide a foundation for machine explanations. Beyond the framework, we will also clarify the notion of machine explainability as a research area, charting its diverse and often confusing literature. To this end, we will outline what, exactly, machine explainability research aims to accomplish.

Finally, we will use all these considerations as a starting point for developing evaluation criteria for good explanations, such as comprehensibility, assessability, and fidelity. Evaluating our framework using these criteria shows that it is a promising approach and augurs to outperform many other explainability approaches that have been developed so far.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Code Listings

List of Abbreviations

AI artificial intelligence. 20, 33, 76, 93, 125, 127, 136, 140, 142, 143, 148, 152, 175, 181, 263, 265, 266

ANN artificial neural network. 18, 19, 129, 158, 161–164, 166, 167, 171, 173, 265, 266

AV autonomous vehicle. 21, 24, 28, 29

CAM class activation map. 161, 164, 165, 172, 174, 201, 203, 211, V, IX

CIA change impact analysis. 158, 159, 172–174

CME causal-mechanical explanation. 116, 117, 126, 181

CNN convolutional neural network. 161, 162, 164–167, 265, V

DL deep learning. 126, 181

DNE deductive-nomological explanation. 116–118, 126, 181, 183, 184, 186

DNN deep neural network. 35, 38, 52, 160, 169, 265

FV feature visualization. 165, 167, 173, 174, 207, 209, V, IX

HLEGAI high level expert group on artificial intelligence. 148

HR human resources. 152

LIME local interpretable model-agnostic explanation. 159–162, 165, 171–175, 193, 210, V

ML machine learning. 17, 19, 22, 23, 44, 45, 51–53, 93, 126, 127, 130, 136, 139, 140, 142, 145–147, 149, 150, 158, 160, 167, 169, 171, 175, 188, 190, 265

NME new mechanist explanation. 116, 117, 126, 181

SDL standard deontic logic. 101–104

TCAV testing with concept activation vectors. 165, 166, 171, 174, 175, 203, V

XAI explainable artificial intelligence. 127

Foreword

In this foreword I would like to briefly talk about two things. First, I would like to say a few words about the origin of this work. After that I come to the acknowledgments.

The Genesis of This Thesis

This thesis partially builds upon several scientific works (i.e., papers and theses) to which I have contributed (other authors depend on the work). Except for one work, all of these are already published. All in all, the thesis builds, thus, upon scientific work conducted over a period of nearly six years (starting from late 2017).

The works in question are as follows (in their chronological order):

- *From Machine Ethics to Machine Explainability and Back*. Published: [60]. *International Symposium on Artificial Intelligence and Mathematics* (ISAIM) 2018. Authors: Kevin Baum, Holger Hermanns, and Timo Speith.
- *Towards a Framework Combining Machine Ethics and Machine Explainability*. Published: [61]. *Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology* (CREST) 2018. Authors: Kevin Baum, Holger Hermanns, and Timo Speith.
- *From Machine Ethics to Machine Explainability and Back – Building up a Framework of Machine Ethics*. Withdrawn after getting a revise and resubmit in a special issue of the *Annals of Mathematics and Artificial Intelligence* (AMAI) 2018. Authors: Kevin Baum, Holger Hermanns, and Timo Speith.
- *A Framework of Verifiable Machine Ethics and Machine Explainability*. Published: [446]. My master's thesis at Saarland University, submitted in August 2018.
- *What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research*. Published: [294] © 2021 Elsevier. *Artificial Intelligence*, vol. 296 (2021). Authors: Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum.
- *Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue*. Published: [105] © 2021 IEEE. *29th IEEE International Requirements Engineering Conference* (RE) 2021. Authors: Larissa Chazette, Wasja Brunotte, and Timo Speith.
- *From Responsibility to Reason-Giving Explainable Artificial Intelligence*. Published: [63]. *Philosophy & Technology*, vol. 35, no. 1 (2022). Authors: Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith.

- *How to Evaluate Explainability? – A Case for Three Criteria*. Published: [448] © 2022 IEEE. *2nd International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *30th IEEE International Requirements Engineering Conference* (RE) 2022. Authors: Timo Speith.

The question that now arises is how these works have influenced this thesis. First, let me comment on the relationship to my master's thesis. Parts of this thesis are a significant revision and extension of my master's thesis. This can be seen most clearly in the second part of this thesis. The idea for the framework, which is built there, already existed in my master's thesis. However, for this work, I have completely rewritten and simplified the formalizations.

My master's thesis was not perfect. Accordingly, one goal of this work was to improve on the mistakes of my master's thesis. Another thing to note is that my master's thesis was in computer science, but this dissertation is in philosophy. Accordingly, I have tried to make the parts that build on ideas from my master's thesis more accessible to a philosophical audience. Finally, the parts of this work in question are based on only a part of my master's thesis, and a large part of the ideas did not find their way into this work at all. Instead, I focused on better embedding the framework, especially with the first part of this work.

Next, I would like to briefly address the influence of the above publications on this work. Written portions or ideas that have been adopted are mostly mine. Even if I have based parts of my work on ideas from these publications that were not originally mine, I have had to significantly adapt these ideas to my argument (this also applies to the ideas that originally came from me). Bringing together the ideas of such distinct works is not possible without significant revision. In addition, all ideas, no matter whom they came from, were significantly thought out and elaborated upon. This is best illustrated by the example of [105]. A significant part of this publication was a systematic literature review, which I also based parts of this work on. However, in order for me to be able to use the results of this review in this thesis, I pretty much completely redid it (more on this in Appendix C).

Taking everything together, this thesis should be sufficiently different from each of the listed works to constitute a new research thesis. Nevertheless, a more detailed breakdown of each section's influences can be found in Appendix E.

Finally, I would like to explicitly state that not all works to which I have contributed have been used as a basis for this thesis. The contents of the following works (listed in their chronological order) have not been used to this end:

- *Explainability as a Non-Functional Requirement*. Published: [277]. RE@Next! Track of the *27th IEEE International Requirements Engineering Conference* (RE) 2019. Authors: Maximilian Köhl, Kevin Baum, Dimitri Bohlender, Markus Langer, Daniel Oster, and Timo Speith.

- *Spare Me the Details: How the Type of Information About Automated Interviews Influences Applicant Reactions*. Published: [293]. *International Journal of Selection*

and Assessment, vol. 29, no. 2 (2021). Authors: Markus Langer, Kevin Baum, Cornelius J. König, Viviane Hähne, Daniel Oster, and Timo Speith.

- *Welcome to the First International Workshop on Requirements Engineering for Explainable Systems (RE4ES)*. Published: [89]. *1st International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *29th IEEE International Requirements Engineering Conference* (RE) 2021. Authors: Wasja Brunotte, Larissa Chazette, Verena Klös, Eric Knauss, Timo Speith, and Andreas Vogelsang.

- *Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives*. Published: [292]. *1st International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *29th IEEE International Requirements Engineering Conference* (RE) 2021. Authors: Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, and Jonas Wahl.

- *On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness*. Published: [264]. *1st International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *29th IEEE International Requirements Engineering Conference* (RE) 2021. Authors: Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz.

- *Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering*. Published: [90]. *28th International Working Conference on Requirements Engineering: Foundation for Software Quality* (REFSQ) 2022. Authors: Wasja Brunotte, Larissa Chazette, Verena Klös, and Timo Speith.

- *A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods*. Published: [447]. *5th ACM Conference on Fairness, Accountability, and Transparency* (FAccT) 2022. Authors: Timo Speith.

- *Explainable Software Systems: From Requirements Analysis to System Evaluation*. Published: [107]. *Requirements Engineering*, vol. 27, no. 4 (2022). Authors: Larissa Chazette, Wasja Brunotte, and Timo Speith. This article is an extension of [105]. While the original article has influenced this work in the way described above, the extension has not.

- *Welcome to the Third International Workshop on Requirements Engineering for Explainable Systems (RE4ES)*. Published: [160]. *3rd International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *31st IEEE International Requirements Engineering Conference* (RE) 2023. Authors: Jakob Droste, Verena Klös, Mersedeh Sadeghi, Maike Schwammberger, and Timo Speith.

- *Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy*. Published: [317]. *3rd International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *31st IEEE International Requirements Engineering Conference* (RE) 2023. Authors: Sara Mann, Barnaby Crook, Lena Kästner, Astrid Schomäcker, and Timo Speith.

- *Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)*. Published: [126]. *3rd International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *31st IEEE International Requirements Engineering Conference* (RE) 2023. Authors: Barnaby Crook, Maximilian Schlüter, and Timo Speith.

- *A New Perspective on Evaluation Methods for Explainable Artificial Intelligence (XAI)*. Published: [449]. *3rd International Workshop on Requirements Engineering for Explainable Systems* (RE4ES), co-located with the *31st IEEE International Requirements Engineering Conference* (RE) 2023. Authors: Timo Speith and Markus Langer.

- All my works that are manuscripts or under peer review at the time of publication of this thesis. For reasons of preserving anonymity, they are not mentioned by name here.

Acknowledgments

First of all, I would like to thank Prof. Dr. Ulrich Nortmann, who has been a mentor and role model to me for many years, providing invaluable support. Furthermore, I am grateful for his extensive feedback on an early version of this thesis. This feedback has significantly improved the work. I am also thankful to Prof. Dr. Marija Slavkovik, who agreed to act as second examiner and to take the time to read and evaluate my work.

I am extremely grateful to Kevin Baum, who first put me in touch with the topic discussed here. Overall, I probably wouldn't have been able to write this thesis if not for the many conversations I've had with him and projects I've tackled with him (including first versions of the proposed framework). For further substantial discussions that made it possible for me to write this dissertation, I would like to thank, in particular, Daniel Oster and Prof. Dr. Markus Langer. In general, I am grateful to all three of them for always encouraging me to continue with this work. At the same time, I am also grateful to them for brightening up my everyday life with lots of humor and little jokes during the entire time I was working on my doctorate.

I would like to thank Prof. Dr. Holger Hermanns for the good cooperation and humorous support. Without him, I would not have had many opportunities, such as many conference visits, and also the opportunity to study computer science. In particular, the many conferences (and the discussions and ideas generated at them) that I have been able to attend with his support have advanced this work immensely.

For further support I would like to thank Prof. Dr. Lena Kästner, who not only gave me the opportunity to finish this thesis without much pressure, but also helped me with words and deeds on the last meters. Additionally, I would like to thank Prof. Holger Sturm (and everyone else involved), who made it possible for me to continue working at the Institute of Philosophy at Saarland University, even though my funding had already run out.

Moreover, for linguistic help, I would especially like to thank A. Kwan, who helped me greatly improve the English of the thesis, while also providing me with food for thought. Also, I would like to thank P. Campbell and Banaby Crook, for the same in smaller parts.

Next, I would like to thank the whole EIS team and the whole AK Explainability, who helped me to better understand and work out my various ideas for this thesis. A big thanks also goes to all my co-authors for the many fruitful discussions. Specifically, I would like to thank Larissa Chazette and Wasja Brunotte for the valuable friendship we developed during our collaboration. No less importantly, I would like to express my gratitude towards everyone from the Institute of Philosophy at Saarland University, especially Oliver Petersen for his philosophical sharpness and Thorsten Helfer, who always had an open ear for a conversation.

I would like to thank the many anonymous reviewers who evaluated the various publications on which this work is based. Their feedback has not only improved the publications in question, but has also indirectly improved this work. I would also like to thank my students for their different perspectives on research and the many things I was able to learn by teaching them.

Many people have indirectly contributed to this work by mentally supporting me and making me who I am today. I will forever be thankful to all of my friends, especially to my friends from Brenkhausen, with whom I have countless good memories; to Angelique Pal Buy, for the many laughs we shared and the encouragement she offered; to Magnus Halbe, without whom I probably would never have gotten my master's degree; to Tobias Sander, for all the tawny owls we looked for together; to Lukas Redemann and Pia Hofmann, for all the good food we enjoyed together; and to all the people I met during my time at the Waldhaus.

Finally, I am deeply grateful to my whole family for their support, patience, and love even when I was very busy with my thesis. In particular, without all the encouragement from my mother, I probably never would have finished this thesis. I am also grateful to my nephews, as they have made me aspire to be a role model (for which a doctorate is an important step in my opinion). Of all thanks one of the biggest goes to my significant other, for simply everything. Without her continuous support, this work would be in a much worse state.

Funding Work on this thesis was funded by the Volkswagen Foundation grants AZ 95143, AZ 9B830, AZ 98509, and AZ 98514 "Explainable Intelligent Systems" (EIS) and by the DFG grant 389792660 as part of TRR 248. The Volkswagen Foundation and the DFG had no role in preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The author declares no other financial interests.

1. Introduction

Artificial computing systems[1] permeate the world in which we live. These systems increasingly infringe upon our lives, and we are rapidly becoming more and more dependent on their functions. A vital question arises from this increasing interaction: How should we constrain machines to behave in a morally acceptable way towards us humans? This question concerns *machine ethics*—the search for formal, unambiguous, algorithmizable, and implementable behavioral constraints on systems so as to compel them to exhibit morally acceptable behavior.

Medical-Care Robot #1

Robots working in hospitals to disinfect rooms or to assist in surgeries have become more commonplace in recent years due to a shortage of care personnel. Over time, medical care robots have intruded into even more sensitive and high-risk applications such as psychological treatment, elder care, and independent resuscitation. If care work is increasingly taken over by robots in the future, there will not only be a debate about whether robotic care is technically feasible, but also morally desirable. [278, 411, 424]

Although some researchers believe that hard-wired ethical constraints on machine behavior are a sufficient precondition for humans to reasonably develop trust in artificial systems, we would like to discuss why this is not the case. Instead, we believe it necessary to complement machine ethics with means by which we can ascertain whether the trust we place in such systems is *justified* and whether they have other desirable properties. After expounding on why this is important, we argue that there is at least one feasible supplement for machine ethics to this end: *machine explainability*—the devising of means by which the actions and decision-making processes of artificial systems can be explained.

Machine explainability thus contributes to machine ethics. This relationship also obtains *vice versa*: machine ethics contributes to machine explainability, as machine explainability can thrive particularly well with a moral system as the basis for generating explanations. Embedded in a moral system, explanations can make reference to moral considerations, thus providing an excellent starting point for calibrated trust (and other desirable properties).

1.1. Machine Ethics

Machine ethics is emerging as a serious area of research, with the first systematic works on it being published recently (e.g., [22, 484]; see also [145] for a brief overview of techniques and challenges). Overall, however, the precise subject of machine ethics is a matter of debate.

James H. Moor pointed out that the term “machine ethics” can be understood quite broadly. According to him, the understanding ranges from the implementation of morally motivated constraints on the behavior of complex and possibly autonomous artificial systems to the

[1] In what follows, we will often just speak of “artificial systems” as an abbreviation for “artificial computing systems”. Furthermore, we will use the term “machine” as a synonym for “artificial computing system”.

implementation of full-fledged moral capacities [346]. On the one hand, the first view is already of great practical importance today because the moral influence exerted through artificial systems, both directly and indirectly, is steadily increasing. On the other hand, the latter view deals with scenarios that remain science fiction and involves discussions about profound philosophical concepts of autonomy, deliberation, and free will.[2]

In this thesis, we will roughly follow the definition set out by Michael Anderson and Susan Leigh Anderson, who understand machine ethics as "concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is [morally] acceptable." [20, p. 15]. Additionally, we will primarily focus on the philosophical dimension of machine ethics, although it is a multidisciplinary field of research [21, 47].

1.2. Machine Explainability

In general, research in machine explainability aims to provide means by which to make various aspects of artificial systems understandable to different audiences [12, 62, 220, 237, 295]. Whether it is the visible behavior, the algorithm on which this behavior is based, or the input required to produce a particular behavior, creating explanations of it is a typical goal of machine explainability research. The superordinate goal of this research, however, is to achieve other desirable properties such as fairness and trustworthiness of artificial systems.

Autonomous Vehicle #1

The functioning of regular cars is often not easy to understand. For example, the software doping that surfaced in the VW diesel emissions scandals revealed plainly that the behavior of complex systems can be extremely difficult—if not practically impossible—for even experts to comprehend [57, 59, 129]. If such problems already occur with normal cars, what will it be like with autonomous cars?

Especially in the context of artificial systems (which often promise positive societal impact), black-box systems whose decisions, predictions, or behavior we cannot accurately explain will not be trusted in the long run. Many applications of artificial systems—for instance, as advisors to politicians and judges—presuppose more than opaque outputs such as numbers (and especially probabilities), at least in the context of liberal democracies. These systems must be auditable, and their results must be justifiable, at least in principle and on request.

Even on the premise that the deployment of some artificial systems is desirable from a moral point of view (thanks to their overall effects), and even if these systems actually behaved as morally well as is logically and conceptually possible (thanks to future advances in machine ethics), as long as people cannot justifiably trust these systems and cannot access the reasons for their decisions, the use of these systems is threatened even where it would be desirable, and cannot be promoted with good conscience in many promising application areas.

[2] We will elaborate on ethics in Section 2.1, and we will further elaborate on these two views in Section 2.2.

However, machine explainability still is a young field of research, and, in particular, formal frameworks supporting machine explanations are scarce (see [121] for a simple one). In this thesis, we want to take the first steps towards a method of performing ethically constrained decision making—machine ethics—in a manner that, in itself, provides a foundation for machine explainability. All in all, it can be summarized that machine ethics and machine explainability should be deeply intertwined.

1.3. The Structure of This Thesis

Arguing for exactly such a deep connection between machine ethics and machine explainability is the overarching goal of this thesis. However, there are also other goals that we will explore. The thesis is hence split into four parts, the first three of which can be read mostly independently. In addition to contributing to the overarching goal of this thesis, each part of it has own subordinate goals.

The first part explores whether machine ethics is a worthwhile endeavor. In particular, we will argue for one specific view of machine ethics: *moral alignment*. Additionally, we will also explore in more detail the connection between machine ethics and machine explainability. Finally, we will discuss what an implementation of machine ethics should look like.

In the second part, we formalize machine ethics in a framework that also enables machine explainability. The framework is based on moral principles and is in line with the kind of implementation we have advocated. Beyond that, we propose an argumentation-based approach to decision making that can be used to generate explanations. By both ensuring machine ethics and enabling machine explainability, the framework forms an important building block for our argument that the two fields are closely connected.

The third part of the thesis deals with machine explainability. Since machine explainability research is still very young, this part of the thesis is mainly about clarification. In particular, we will extract the goals of machine explainability from the literature and summarize them in a model. These goals are, again, linked to machine ethics, and constitute the third ingredient in our argument for the close connection between machine ethics and machine explainability. To conclude the presentation of machine explainability, we will present some approaches that aim to provide explainability and devise quality criteria for them.

The fourth and final part connects the first three parts. Taking into account our views on machine ethics and machine explainability, we will apply the quality criteria we have devised for explainability approaches to our framework and argue for its appropriateness and comparability to other approaches.

These descriptions of the parts should suffice as an introduction to the thesis. For the interested reader, a more detailed description follows.

The next section of this thesis (Section 2) will position machine ethics, via a broad overview of ethics, as a branch of applied ethics. Additionally, we will present the main research strands of machine ethics, and we will discuss some recent approaches to it.

In Section 3, we will examine machine ethics in more detail. In particular, we will discuss whether it is worthwhile to engage in research in this area, or whether such research could bring more disadvantages than advantages. By carving out our own view on machine ethics, we will show that research is indeed worthwhile, and especially so if it is supported by research in machine explainability.

The subsequent section (Section 4) will build a bridge to the framework we envision by providing initial arguments on what such a framework should look like. Additionally, this section will include machine ethics research itself by discussing the (dis)advantages of programming an artificial system with one of the traditional normative theories.

We begin the second part of this thesis by developing our framework for machine ethics (see Section 5). This framework is based on moral principles, which are intended to play an essential role in decision making.

The second section (Section 6) will then outline the detailed decision-making process. This process relies on arguments in a graph-based approach to provide a fruitful basis for creating explanations.

In Section 7, the second part of this thesis finishes with some additional thoughts and ideas concerning the framework. In particular, we will discuss approaches to modeling the moral principles on which the framework is built. Moreover we outline an alternative formalization for the framework that builds on STIT logic.

Section 8 brings us to machine explainability. Here, we give an overview of the field of explanation research to demonstrate its peculiarities. This section will also introduce our model of the processes in machine explainability and how they relate to each other.

Further deepening the model, Section 9 links back to machine ethics. Via an extensive literature review of over 200 papers, we highlight the goals pursued by research in machine explainability, many of which have moral components. For one of these goals, we explore in more detail how machine explainability should contribute to it.

The final section of the third part of this thesis (Section 10) discusses explainability approaches. First, we present some exemplary approaches, after which we turn to the evaluation of such approaches. Taking into account, among other things, the goals of machine explainability, we develop and justify our own quality criteria for explanations. We apply the criteria to the discussed approaches, most of which do not pass the test.

Finally, the thesis ends in a discussion of why our framework can, in principle, pass the test (Section 11).

Part I.
Machine Ethics

2. Charting the Field of (Machine) Ethics

In order to discuss machine ethics, we first have to clarify the subject matter of this research discipline. This section is concerned with doing so. As machine ethics is a branch of ethics, we will begin with the subject matter of ethics more broadly. Subsequently, we will zero in on machine ethics and its goal(s) and varieties. Finally, we will discuss some contemporary approaches to machine ethics to convey a better idea of this research discipline.

2.1. What is Ethics?

When speaking of ethics, one usually also thinks of morals. Commonly, the terms "ethics" and "morals" are used interchangeably. In philosophy, however, there is a clear distinction between them, which we will also follow in this thesis. Ethics (or, as it is also called, *moral philosophy*) is the field of research concerned with morals. As this definition may not be very illuminating, let us give another description. Roughly, ethics involves systematizing, defending, and recommending concepts of right and wrong behavior [174]. Morals, on the other hand, are these concepts of right and wrong behavior that ethics is concerned with [159].

2.1.1. The Three Main Branches of Ethics

Nowadays, ethics is usually divided into three main branches: *metaethics*, *normative ethics*, and *applied ethics* [149, 174]. Let us briefly illuminate each of these branches.

Metaethics The ancient Greek term "meta" means "after" or "beyond", and, consequently, the term "metaethics" implies a distanced, or bird's-eye, view of the whole endeavor of ethics [174]. Metaethics is the attempt to understand the metaphysical, epistemological, semantic, and psychological presuppositions and commitments of moral thought, talk, and practice [413]. To put it briefly, while applied ethics and normative theory focus on what practices are moral, metaethics focuses on what morality itself is [143]. In this way, metaethics can be understood as a highly abstract way of philosophical thinking about morality. For this reason, metaethics is occasionally referred to as "second-order" moral theorizing to distinguish it from the "first-order" level of normative theory [143].

Compared to normative ethics and applied ethics, metaethics is the least well-defined subfield of moral philosophy [174]. Two issues, however, are prominent: (1) metaphysical issues concerning the question of whether morality exists independently of human beings, and (2) psychological issues concerning the underlying mental basis of our moral judgments and conduct [174]. Some questions belonging to the first group of issues are "Is morality a question of taste rather than of truth?" and "Are moral standards culturally relative?" [413]. Questions belonging to the second group of issues are "How might moral facts be related to other facts (about psychology, happiness, human conventions, . . .)?" and "How do we learn of moral facts, if there are any?" [413].

Normative Ethics Normative ethics can be seen as much more practical than metaethics. This branch of ethics deals with the question of what makes an act right or wrong. Traditionally, the formulation of normative theories (i.e., theories about what moral norms govern right and wrong conduct) falls within normative ethics. One could say that the goal of normative ethics is to find the ultimate test for proper behavior, and the fundamental assumption in normative ethics is that there is only one ultimate criterion for moral conduct. [174]

Example #1

A classic example of a normative principle is the *Golden Rule*: we should do to others what we would want others to do to us. The Golden Rule is an example of a normative theory that establishes a single principle by which we judge all actions. [174]

There are many normative theories, and we will present the three major families of theories that are commonly discussed in normative ethics in the next subsection (Section 2.1.2).

Applied Ethics Applied ethics is concerned with examining specific, controversial moral issues. Such controversial issues include abortion, the death penalty, animal rights, and euthanasia [174]. Applied ethics attempts to answer the question of whether it is morally justifiable to introduce or do these things.

In recent years, applied ethical issues have been subdivided into subject-specific groups such as medical ethics and business ethics. The reasons for this development are manifold. First, what applied ethics investigates is quite diverse. Second, work in this field requires considerable empirical knowledge. Finally, the pursuit of applied ethics has historically been done by looking at different kinds of human practices. Taking all of this together, it only makes sense that there will be many different types of applied ethical research, such that an expert working in one type will not have much to say in another. [149]

In general, two features are required to constitute an "applied ethical issue". First, the issue must be controversial in the sense that there are significant groups of people both for and against it (see Example #2). [174]

Example #2

Take the issue of killing out of base motives (i.e., murder). This issue is not an applied ethical issue since everyone agrees that murder is fundamentally immoral. In contrast, the issue of abortion is an applied ethical issue since there are significant groups of people both for and against abortion.

Second, to be considered an applied ethics issue, the issue must be distinctly moral [174]. Take the example of raising wealth taxes. Although this issue is controversial and has an essential impact on society, it is not distinctly moral.

Here, it is important to distinguish social policy from moral issues. The aim of social policy is to increase the efficiency with which a given society runs. In contrast, moral issues concern more universally obligatory practices (as mentioned above). While the two often overlap (e.g., when condemning murder), there are many cases in which they do not (see Example #3).

Example #3

Sexual promiscuity is often seen as immoral, whereas it is commonly not regulated by social policies. On the other hand, in some neighborhoods, yard sales are forbidden by social policies, whereas there is nothing really immoral about yard sales if the other residents are not offended by them in any relevant way. [174]

Interplay Between The Branches The three branches of ethics overlap, and the demarcations between them are often blurry. Additionally, there is a considerable interplay between them. Let us take applied ethics as an example of such interplay. In order to judge the moral status of a particular controversial practice (say, abortion), results from normative ethics and metaethics must be employed. Metaethics can state the moment from which a fertilized oocyte starts to be morally relevant, and normative ethics delivers theories that give the criteria by which the act of abortion itself must be judged.

In philosophy, machine ethics can be seen as a branch of applied ethics. As such, it also interplays with normative ethics and metaethics, as we will examine in more detail later on. For now, a brief example will suffice to indicate the nature of this interplay: when it comes to the connection with metaethics, for instance, the debates in machine ethics revolve around questions similar to those in abortion, including, in particular, the question of when an entity qualifies as worthy of moral consideration (i.e., when it qualifies as a *moral patient*).

2.1.2. The Three Major Families of Normative Theories

As the interplay of machine ethics and normative ethics is crucial in what follows, we want to discuss normative ethics before we come to machine ethics. Currently, there are three major families of normative theories: *deontological theories*, *consequentialist theories*, and *virtue theories*. These theories are distinguished mainly by their respective foci. When judging the moral status of an action, deontological theories commonly focus on the *action* itself, consequentialist theories on the *consequences* of the action, and virtue ethics on the (virtues of the) *agent* performing the action (see Figure 1 for a visualization).

These differences in focus do not mean that only consequentialists take consequences into account. All three families of normative theories can make room for consequences, rules, and virtues. What differentiates these families of theories is the centrality of one of these concepts within the family. For virtue theories, for instance, virtues and vices will be foundational, and other normative notions will be grounded in them. [249, 266]

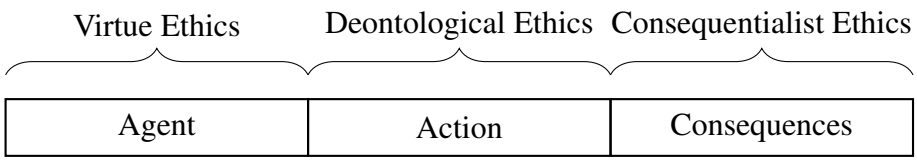


Figure 1: The three major families of normative theories and their respective foci.

Deontological Ethics The word "deontology" derives from ancient Greek and means "study of duty" or "science of duty" [7, 8]. Deontological theories are sometimes also called *duty theories* (as the name suggests) or *non-consequentialist theories*. The latter name is meant to contrast them against consequentialist theories, and to highlight that, in deontological theories, the consequences of an action do not primarily influence its normative status [174]. Deontological theories focus on actions. More precisely, deontological theories base the moral status of an action on specific, basal principles of obligation [174]. Such principles are, for instance, not to commit homicide, or to care for one's family.

The arguably most famous branch in the family of deontological theories is *Kantianism* (see [261] for its source). With the categorical imperative, Immanuel Kant developed a deontological principle that requires agents to perform only those actions that follow maxims (i.e., subjective rules or policies of action) that can be universally followed without contradiction [58, 255, 398]. In its original formulation, the categorical imperative reads as follows: "act only on that maxim through which you can at the same time will that it should become a universal law" [262, p. 88]. There are several formulations of the categorical imperative that Kant considered equivalent, although many researchers dispute this equivalence claim [255].

Kant's theory is often criticized for its inability to accommodate crucial nuances regarding certain moral dilemmas (see Example #4). Many people find this affair unsatisfactory. [7, 8]

Example #4

One example of where Kantianism fails concerns a person sheltering a Jewish person from Nazis. The person has a duty to protect the Jewish person, but if the Nazis knock on their door and ask whether they shelter a Jew they also have a duty not to lie. In this case, the only relevant action still allowed, according to Kantianism, is to remain silent.

For this reason, ethicists developed more fine-grained deontological theories. One such theory is that of William D. Ross (see [402, 403] for its source). Ross distinguished seven so-called *prima-facie* duties[3] (viz., fidelity, reparation, gratitude, non-injury, beneficence, self-improvement, and justice, v. [174]) that determine the moral status of an action. Although

[3]Roughly, the term "prima facie" means something like "at first glance". However, there is some debate as to whether the term "prima facie" is appropriate for what Ross is trying to express in his theory. Shelly Kagan, for example, wrote: "It may be helpful to note explicitly that in distinguishing between pro tanto and prima facie reasons I depart from the unfortunate terminology proposed by Ross, which has invited confusion and misunderstanding. I take it that [...] it is actually pro tanto reasons that Ross has in mind in his discussion of what he calls prima facie duties." [258, p. 17]. We will come back to the term "pro tanto" later.

his theory, arguably, can deal with scenarios like the one just described, it has other problems. Most prominently, it is not always obvious what to do in cases of competing duties [8, 23, 24].

Consequentialist Ethics As its name suggests, consequentialist ethics focuses on consequences. More precisely, consequentialism is the view that the normative properties of an action depend solely on its actual or expected consequences. For this reason, consequentialist theories are sometimes called *teleological theories*, from the ancient Greek word for "end" (i.e., "telos"), as the result of an action is the sole determining factor of its morality. [174]

The consequentialist picture is driven by maximizing value and whose (moral) value (and disvalue) is judged are states of affairs, how things are. Hence, what makes an action right (or wrong) is what the action changes in the world. An action that brings about (or promises to bring about) more value than disvalue is better than an action that brings about (or promises to bring about) more disvalue than value. Often, then, the action that brings about (or promises to bring about) the highest net value is not seen only as the best action, but as the *right* action—or as one of the right actions, if there are multiple such actions possible. [446]

Particularly well-known theories that belong to the consequentialist family are *utilitarian* ones. What distinguishes utilitarian theories from other consequentialist theories (such as *ethical egoism* and *altruism*) is that they commonly consider the interests of all humans (or, more generally, those of all moral patients) equally [204]. However, proponents of utilitarian theories disagree on other issues. For instance, they disagree about whether the actual or the expected consequences of an action count. Furthermore, they disagree on whether it should be individual acts that count (*act utilitarianism*) or whether agents should, more generally, follow rules adherence to which promises maximal value (*rule utilitarianism*). Lastly, utilitarians disagree on whether pleasure and pain (hedonistic theories), preference satisfaction and frustration (preference theories), or other things (e.g., objective goods) count as value. [435]

A famous member in the family of utilitarian theories is *classical utilitarianism*. Classical utilitarianism is a form of utilitarianism in which the actual consequences of an action count. Furthermore, what counts as value and disvalue for classical utilitarians are pleasure and pain, respectively. Historically, the central proponents of (classical) utilitarianism include Jeremy Bentham [68] and John Stuart Mill [338]. More recent advocates of (different types of) utilitarianism include Richard M. Hare [211] and Peter Singer [433]. Overall, Bykvist offers a systematic discussion of utilitarianism (in [94]). [58]

In contrast to deontological theories, consequentialist theories are not vulnerable to moral dilemmas. This invulnerability stems from the fact that consequentialist theories can always recommend a course of action. If several actions maximize the value, it does not matter which one to take; one just has to take one. However, consequentialist theories are subject to a host of other objections. For example, consequentialist theories are often criticized for allowing actions that are commonly seen as fundamentally immoral. To say it bluntly, if the end justifies the means, the means can get pretty ugly [197, p. 52].

Virtue Ethics As we have seen, many philosophers believe that morality consists of following well-defined rules of conduct or adhering to certain duties (as deontologists advocate), or of acting to bring about good consequences (as consequentialists advocate). Presumably, one must learn these rules (or assess the consequences) and ensure that each of one's actions is in accordance with the rules (or brings about the best consequences). [174]

Example #5

Suppose it is evident that someone in need should be helped. A utilitarian will point out that the consequences of doing so will maximize well-being, and a deontologist will say that the agent will act according to a moral rule (e.g., the Golden Rule). [249]

Virtue ethicists, however, place less emphasis on learning rules or assessing consequences, and instead stress the importance of developing ideal character traits such as charity, temperance, or justice [174]. These ideal traits that one possesses, or aims to possess, are the *virtues* [184]. Likewise, virtue theorists believe that we should avoid acquiring bad character traits (i.e., *vices*) such as cowardice, injustice, or vanity [174].

Virtue ethicists commonly advocate that right actions will tend to come from a virtuous person. In this way, virtues are often seen as sets of stable dispositions to act in certain ways responsive to characteristics of one's environment [184]. For example, once a person has acquired charity, they will, then, habitually act charitably [42].

Example #6

In the above case (Example #5), a virtue ethicist will point out that helping the person would be charitable or benevolent, and something that a virtuous person would do. [249]

Character traits derive from natural internal tendencies, but they must be nurtured. In this line of thought, virtue ethicists emphasize moral education since character traits are developed in one's youth. Adults, therefore, are responsible for instilling virtues in the young. [174]

Historically, virtue theory is one of the oldest normative traditions in Western philosophy, having its roots in ancient Greek civilization with proponents like Plato and Aristotle. Arguably, versions of virtue ethics are present also in ethical traditions of Eastern philosophy, such as Confucianism, Daoism, and Buddhism (see [269, 429] for arguments in this direction).

In the Aristotelian view, virtues contribute to, or even constitute, what he called *eudaimonia* (meaning something like "flourishing"). For this reason, his theory is called *Eudaimonism*. Plato emphasized four virtues, which came to be called the *cardinal virtues*: wisdom, courage, temperance, and justice [174]. While the Platonic version of virtue ethics was especially prevalent in the Middle Ages as part of Scholasticism, versions standing in the Aristotelian tradition are increasingly popular nowadays. Proponents of such modern forms of virtue ethics are G. E. M. Anscombe [31], Philippa Foot [176], and Rosalind Hursthouse [248]. [42]

Some virtue theories are criticized because they do not guide action (unlike the other two families of normative theories) and depend on luck: not everyone has the chance to make real friends, so whether one can flourish as a person seems to depend on luck. For some ethicists, these factors disqualify virtue ethics from being a proper normative theory. [42, 249]

2.2. What is Machine Ethics?

With all of the above in mind, we can come back to machine ethics to shed more light on it. As already stated in Section 1.1, we will roughly follow the characterization set out by Michael Anderson and Susan Leigh Anderson, who understand machine ethics as "concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is [morally] acceptable." [20, p. 15].

Arguably, their definition is rather broad and also involves what is traditionally known as "safety engineering". Indeed, many machines have been engineered with thoughts of morality or beneficence for humans in mind before machine ethics emerged as a research discipline. Such systems, commonly, have been restricted in their behavior for certain reasons (e.g., because their exploitation can lead to bad consequences). Against this background, the question arises as to what extent machine ethics differs from or augments safety engineering.

Example #7

Take cash machines of banks as an example. In general, these systems are designed so that they are difficult to exploit. In particular, they usually make it hard to get access to other people's bank card credentials (and, thus, to other people's bank accounts). [100]

There is one central aspect that distinguishes machine ethics from safety engineering: the extent of system adaptability. Artificial systems are deployed in more and more contexts, and they operate increasingly autonomously. In general, traditional safety engineering methods do not suffice when machines must operate in changing environments: under changing conditions, they need flexibility [11, 346]. Furthermore, these methods reach their limits when systems have to operate under high uncertainty [100]. Finally, safety engineering does not arrive at satisfactory solutions when there are conflicting courses of action to choose between [100].

2.2.1. The Goal(s) of Machine Ethics

When traditional safety engineering reaches its limits, machine ethics comes to the rescue. Its main idea is that equipping machines with capacity for moral reasoning can help to make them more moral (from the viewpoint of humans) across a wide variety of contexts. This is one of the goals of machine ethics. However, there are also other goals, and it is possible to distinguish roughly two strands of research in machine ethics. These two strands of research correspond approximately to the distinction made by Moor (see Section 1.1).

We have just mentioned the goal that the first strand of research pursues. Here, the leading question is how to equip systems with some representation of moral values and a decision procedure that factors them in. Thus, this strand of research is concerned with the *moral behavior* of artificial systems. The second strand of research, in contrast, is concerned with the *moral status* of machines. Here, the goal is to clarify whether machines actually have, could (eventually) have, or even should have moral status. Depending on the answers to these questions, we would have to change our behavior towards artificial systems significantly.

In this thesis, the first—behavioral—strand of machine ethics research takes a pivotal role. However, before we give an overview of contemporary approaches to making machines behave in more morally acceptable ways (in Section 2.2.2), we would like to shed light on several conceptions of how the goals in both research strands are supposed to be achieved.

The First Strand of Machine Ethics Research: Moral Behavior A modest concept of the goal that the first strand of research in machine ethics pursues is to achieve what has been called "ethical alignment" [100]. According to this concept, the goal of machine ethics is to make the behavior of machines more morally desirable from the perspective of humans, even if only by a little bit [100, 377, 378]. A slightly morally desirable machine deployed is better than one that is not morally desirable but deployed nevertheless. We note that "*ethical* alignment" is not the appropriate term for this goal, given the common distinction between ethics and morality. Therefore, we will speak of "*moral* alignment" in what follows.

This concept stands in contrast against several others in machine ethics. First, there is the view that machine ethics must lead to a specific (and maybe unattainable) level of morally desirable behavior; that machines that do not exhibit a certain level of morals should not get deployed at all [10]. While this may be true, the spread of artificial systems is not likely to lessen soon [23]. Unfortunately, considerations of morality are only slowly proliferating into the debate, and systems that should not get deployed are brought into use nonetheless. For this reason, at least for now, moral alignment is a feasible goal of machine ethics that promises to alleviate the most pressing problems.

Second, there is the view that contemporary machines cannot "act" in a relevant sense of acting and, thus, cannot act morally [10]. This view is deeply connected with the discussions about (artificial) agency and, in particular, with (artificial) moral agency. If machines do not qualify as moral agents, it does not make sense to get them to act morally. Predicates of morality simply cannot be attributed to them. Consequently, if a system is supposed to act morally, it must qualify as a moral agent (i.e., as an entity that can act based on morals). Accordingly, people who take this view of machine ethics advocate that the goal of machine ethics should be to create machines that qualify as moral agents by equipping them with capacities traditionally associated with (moral) agency, such as sentience, sapience, and autonomy [346]. Their moral behavior, then, follows from these capacities.

We will come to the latter view later on in more detail (in Section 4.1.1), but for now it should suffice to state that, at least in our opinion, it misses the point of machine ethics. Machine ethics is not about equipping machines with the capacities to be able to act morally. It is about equipping them with the capacity to factor moral considerations into their decision-making processes. Artificial systems do not need to be aware or conscious of the fact that such considerations constrain them. What counts is that the machines are behaving in a morally desirable way from an outside perspective (i.e., from a human perspective).

In this thesis, we will argue that moral alignment is sufficient to harness the advantages that machine ethics promises to bring while avoiding its potential pitfalls and drawbacks. Before we do so, however, let us first contemplate the second strand of research in machine ethics.

The Second Strand of Machine Ethics Research: Moral Status We can distinguish several approaches with respect to machines' moral status. We have discussed the motivation behind the first approach above: the belief that many approaches in the first strand of research do not suffice to reach their goal. Even if we could implement morally motivated restrictions into systems, this would not be sufficient to reach adequate moral standards [10].

People who share this belief argue that for artificial systems to exhibit a satisfactory level of morally desirable behavior, they must be equipped with capacities needed for truly moral decision making (i.e., they must become moral agents) [10, 11, 346]. Such capacities are, for example, sentience, sapience, and autonomy. The moral status of machines, then, stems from them having these capacities: moral agency commonly implies moral patiency [224].

A second approach stems from the belief that, as artificial systems are equipped with increasingly sophisticated capacities for moral reasoning, they inevitably acquire a moral status at some point [100, 140]. In this context, the moral status of the machines stems from their increasing sophistication and is complicated by our current inability to measure whether an entity has capacities like sentience or sapience to a level that qualifies it as a moral patient [202]. We simply do not know enough about concepts like consciousness to reliably attribute them [202].

The third approach we want to mention originates from the belief that traditional criteria for moral patiency are somehow inadequate. People who share this belief often argue that an entity's moral status depends on its relational status in a society and not on other criteria (like having certain capacities) [120, 253]. Other people with this belief argue that we should grant machines a moral status to be safe rather than sorry. For these people, it could be that, unbeknownst to us, machines already have capacities that qualify them as moral patients, but we do not notice so, either because our criteria are wrong or because they are too vague [119].

Finally, some people advocate that artificial systems may not have moral status (yet), but we should treat them as if they did have one so as not to morally corrupt ourselves [136, 439].

In general, most researchers deny machines a moral status [100]. We will briefly discuss the topic of moral patiency with regard to artificial systems in Section 3.1.2 and Section 3.1.3.

Nevertheless, the discussion about moral patiency already pointed to one aspect that needs further illumination: there seem to be potential demerits connected with research in machine ethics. For this reason, we will discuss the potential merits and demerits connected with machine ethics in the next section (Section 3.1). However, before we come to that, we will discuss some contemporary approaches to machine ethics in the following subsection.

2.2.2. Approaches to Machine Ethics

There is already a wide variety of approaches to machine ethics. Brundage [88] distinguished five classes of approaches to it. The first three classes are based on a previous distinction made by Allen, Smit, and Wallach [9]: *top-down*, *bottom-up*, and *hybrid*. Furthermore, Brundage discusses *psychological* approaches to machine ethics and approaches to ethical artificial general intelligence. While this last class of approaches is not relevant for our present purposes (as we are not directly concerned with artificial general intelligence), we will discuss some examples that belong to the other four classes in what follows.

Top-Down Approaches Top-down approaches usually try to implement one or more well-known members of one of the traditional families of normative theories as directly as possible in a machine. We will comment more on top-down approaches later (in Section 4.1).

Deontological Approaches On a general level, Powers discusses various possibilities for a deontological ethical machine. Inspired by Kant's categorical imperative, Powers examines three plausible accounts of deontic logic for implementing a rule-based moral machine in the Kantian tradition: one based on mere consistency, one based on commonsense practical reasoning, and one based on coherency. Unfortunately, all three accounts face serious challenges (e.g., excessive specificity, lack of semi-decidability, or lack of priority of maxims). Nevertheless, Powers believes that these challenges can, eventually, be overcome in a way that might also help resolve some of the human ethical challenges. [377]

In an early work, Anderson, Anderson, and Amen propose an approach to machine ethics based on William D. Ross' prima-facie duties. After noting that Ross did not elaborate on how to weigh his seven duties, they contend that a method inspired by Rawls' reflective equilibrium could help solve this problem. Such a method would involve tweaking the relative weightings of each of the seven duties until the judgment of particular cases conforms with our intuitions. Since performing such weighting manually would quickly overwhelm humans, it lends itself to implementation in machines. In this line of thought, they developed a prototype ethical advisor system: *W.D.* [23]

Leben proposes a Rawlsian approach to develop morally aligned autonomous vehicles. The basic idea is to calculate which action each agent would agree to in a dilemma if they were in an original position behind the veil of ignorance. Leben's approach is guided by the *maximin* criterion (i.e., that the worst-off person should be made as well-off as possible). To elaborate

on his idea: in an original position, in which the agents do not know (because they are behind the veil of ignorance) whether they are potential car occupants or pedestrians with whom the car might collide, even self-interested individuals would agree that it is most beneficial to make the worst-off person as well-off as possible. Leben argues that his approach leads to unambiguous solutions in most cases and is better than a comparable utilitarian approach. [300]

Consequentialist Approaches In the same work where they propose an approach to machine ethics based on William D. Ross' prima-facie duties, Anderson, Anderson, and Amen also propose a utilitarian approach. More specifically, they propose an approach based on *hedonistic act utilitarianism*. Anderson et al. argue that such an approach lends itself to being implemented in a machine, as it is *consistent* (i.e., for a given set of inputs, the implementation will always arrive at the same output), *complete* (i.e., for all valid inputs, the implementation will arrive at valid outputs), and *practical* (i.e., the implementation will reach a conclusion in a reasonable time). Similar to the other approach they present in their paper, they also develop a prototype ethical advisor system for this approach: *Jeremy*. [23]

Although he does not discuss a specific approach to machine ethics, Grau examines whether robots should decide based on utilitarian criteria. He finds that some of the traditional objections against utilitarianism do not hold if the acting entity is a robot that does not qualify as a moral patient (we will look more closely at his findings in Section 4.2.1). Grau argues that we should prevent robots from ever attaining such a status. Nevertheless, he also finds that other objections against utilitarianism still apply, and concludes that robots should not act in a utilitarian way if humans are concerned. However, Grau does not preclude robots from acting in such a way against fellow robots, as long as these do not qualify as moral patients. [197]

Whereas Anderson et al. and Grau focus on *act utilitarianism*, Bauer proposes a *two-level utilitarianism*: rule utilitarianism is applied first, and only then act utilitarianism. Bauer motivates his approach by comparing it to a hybrid approach proposed by Howard and Muntean based on virtue ethics (we will summarize this approach below), and argues that he harnesses the advantages of that approach while avoiding its drawbacks. In general, Bauer's approach is to program an artificial system with a set of (primarily domain-specific) moral rules that govern its overall moral behavior. In cases where no rules apply or where rules are conflicting, an act-utilitarian calculus comes into play to determine the system's action. Furthermore, Bauer elaborates why his two-level approach is superior to one-level approaches: it closely captures our human moral lives and ways of thinking. [58]

Bottom-Up Approaches Typically, bottom-up approaches use machine learning (ML) techniques to find some manner of morally aligned decision making. The machine is presented with various situations that require moral judgment. Based on what ethicists think of these

situations, the machine tries to extrapolate how it should behave in general. Bottom-up approaches to machine ethics depend heavily on the training data used, so the ethical values of the humans training such systems cannot be separated from the computational framework being trained [88].

Guarim trained artificial neural networks (ANNs) on various moral judgments. His goal was to model these judgments with the networks in order to gain new insights into the generalism–particularism debate (i.e., whether moral judgments are always based on immutable rules or principles, or whether they depend on context). Through his experiments, Guarim argues for a middle ground between generalism and particularism: although the ANNs learn from particular cases, they exhibit some sort of general rule by which they classify new situations. Similarly, some situations deviated from the norm and did not lend themselves directly to being interpreted as a rule. [200]

McLaren proposes case-based systems to help humans make moral decisions. His early Truth-Teller system compares different situations and highlights morally significant differences between them [38, 330]. His later SIROCCO system emulates how an ethical review board within a professional engineering organization decides cases by referring to, and balancing between, ethical codes of conduct and past cases [327]. McLaren suggests that his two systems can work together with a high degree of synergy: while SIROCCO only refers to similar cases without specifying where the similarities and differences lie, Truth-Teller only compares cases without referring to specific principles. [328, 329]

Hybrid Approaches As the name suggests, hybrid approaches combine top-down and bottom-up approaches. Initially, the artificial system is programmed with reference to a normative theory, leaving some variables open. For instance, the theory's overall decision-making procedure is mimicked. Afterward, the variables previously left open are "filled in" by learning from examples. Take consequentialism as an example: one possible factor that the machine can learn is the (moral) weight it attaches to certain consequences.

One of the first proposals that can be considered a hybrid approach is Cloos' *Utilibot*. Cloos discusses that initial attempts to create moral robots are likely to be lacking. Nevertheless, he holds that subsequent improvements will ameliorate initial faults. With this in mind, he suggests a utilitarianism-based approach to building moral machines. His Utilibot consists of four modules: the first one models the user and their health, the second models the environment, the third is responsible for assigning utilities to certain states of the user and the environment, and the last module plans the robot's actions. The utilities that the Utilibot assigns are learned based on the impact of specific actions on the user's health (making Cloos' approach a hybrid one). Furthermore, Cloos suggests that the Utilibot should be developed through three generations: in the first generation, the value that the Utilibot tracks is physiological health; in the second generation, hedonic well-being is also factored in; finally, the third generation is supposed to track what is commonly seen as "happiness". [117]

Building on their earlier results, Anderson and Anderson further develop their approach to Ross' prima-facie duties. To recap, their approach refines Ross' theory of prima-facie duties with ideas from Rawls' reflective equilibrium to arrive at an appropriate weighting of duties. Using ML in their approach, Anderson and Anderson have discovered a decision procedure for weighing prima-facie duties in the domain of medical ethics that had not been articulated before, yet conforms to expert opinion. [28]

Virtue ethics lends itself to implementation as a hybrid approach, as virtue ethics generally emphasizes learning as an essential part of becoming virtuous. Accordingly, virtue ethics takes a prominent role when it comes to hybrid approaches to machine ethics.

In this line of thought, Howard and Muntean draw several analogies between human and machine learning that showcase the plausibility of a virtue-based, morally aligned, artificial system. Their focus is on acquiring (moral) skills and behavioral dispositions, which they understand as virtues. Although they do not elaborate on the robot's architecture, they claim that these skills and dispositions may be acquired using ANNs and evolutionary computing. [243, 244]

Another example is Wiltshire's heroism approach. For him, an ideal moral agent forgoes self-interest and may even harm itself for the greater good, just like heroes do. In contrast to Howard and Muntean, Wiltshire even sketches a decision-making architecture for a robot that acts in such a way. Furthermore, he suggests that the robot may initially learn how to act based on examples from film and fiction. To illustrate that this is a plausible idea, he cites some examples of robots in movies that have acted heroically. [496]

Integrating ideas behind the above approaches, Pontier and Hoorn developed the *Moral Coppelia*. They take dual-process theories of moral judgment seriously and combine top-down knowledge in the form of moral duties (viz., autonomy, non-maleficence, and beneficence) with bottom-up structures. Fulfilling these three (sometimes conflicting) duties serves as the goal that the system should reach. Similar to some of the other approaches introduced in this section, Pontier and Hoorn evaluated their approach with simulation experiments that showed congruence with expert judgment. [374]

Psychological Approaches Unlike the first three kinds of approaches, psychological approaches do not attempt to emulate moral behavior directly based on a normative theory or expert opinion. Instead, they try to mimic the psychological decision-making process that humans employ when coming to a (morally impactful) decision. Although one may argue that normative theories model human psychology at least in part, as moral views depend, to some extent, on intuitive or moral processes, psychological approaches place particular emphasis on psychological processes, and draw specifically on insights from cognitive science and neuroscience (which the other approaches do not) [88].

Dehghani et al. developed *MoralDM* to capture psychological results concerning how humans make moral decisions. Overall, this results in an approach with two central components:

a first-principles reason module, and an analogical reasoning module. The first-principles reasoning module integrates both utilitarian and deontological calculations. Specifically, utilitarian calculations are used as long as no so-called "sacred" values (i.e., values with a very high cultural or religious worth, e.g., human lives) are at stake. Otherwise, deontological calculations are preferred. The analogical reasoning module captures the psychological finding that humans tend to think in terms of analogies and comparisons. This module links a given situation that is to be morally judged with past situations of which moral judgment is available. If the situations have a high degree of similarity, the outcome of the analogical reasoning module is preferred to that of the first-principles module. Dehghani et al. have investigated the quality of MoralMD in three studies, finding that its outputs were consistent with expert opinion. [142]

Gomila and Amengual base their ideas on the role of emotions in moral cognition. To account for Frijda's finding that emotions comprise five distinct components (an evaluation of a perceived situation, a qualitative sensation, a type of psychological arousal, an expressive component, and a behavioral disposition, v. [181]), they propose a hierarchically organized system based on behavior-based artificial intelligence (AI). In this system, the agent's primary functions are at the top of the hierarchy, while progressively more specialized skills are situated further down. Emotions are linked closer to the top of the hierarchy. Thus, the higher the implications of a particular event in the hierarchy, the more emotionally arousing it is. According to Gomila and Amengual, such organization has the additional benefit of enabling global appraisal of oneself in relation to other entities. This benefit is particularly important when it comes to motivations such as socialization and attachment, and allows for the generation of emotions such as shame and remorse. [191]

This last aspect of empathizing with others is also essential for Bello and Bringsjord's approach. Their approach takes construal level theory as a starting point to construct what they call a "computational model of mindreading". Construal level theory is a psychological framework for how humans think about events outside their immediate periphery, for instance, by imagining hypotheticals, perspectives of other agents, and counterfactual alternatives. Bello and Bringsjord's basic assumption is that moral machines must have something like a moral commonsense, and that the best way to emulate such a moral commonsense is by mimicking folk intuitions about morals. Based on psychological experiments showing that how people reason about an agent's behavior differs according to the psychological distance from which it is viewed (something that construal level theory can account for), they encode psychological distance as a cost on logical constraints that govern the agent's behavior. [67]

Now that we have thoroughly introduced and discussed machine ethics as a field of research, it is important to determine—in view of the possible disadvantages already mentioned—whether it makes sense to research this field. The goal of the next section is to do just that.

3. From Machine Ethics to Machine Explainability

In the literature, there are various pro-tanto reasons[4] for and against pursuing research in machine ethics. In this section, we will argue that the reasons for pursuing the research discipline of machine ethics do, indeed, outweigh the reasons against doing so. Furthermore, we will argue that machine explainability can serve as a catalyst for machine ethics, profoundly contributing to this research discipline in numerous ways. Overall, this type of connection gives a crucial reason to also pursue research on machine explainability.

3.1. Is Machine Ethics Worthwhile?

Here, we will examine the reasons for and against pursuing machine ethics without factoring in machine explainability. The goal is to show that the reasons in favor of outweigh the reasons against this pursuit, largely because these reasons against it can be easily refuted.

3.1.1. Reasons for Machine Ethics

We distinguish two families of reasons for pursuing machine ethics research. In what follows, we will call these reasons "motivations" to demarcate them more clearly from reasons against machine ethics research (which we will call "risks").

Motivation 1: The Well-Being of Humankind The first family of reasons for pursuing machine ethics research is concerned with the well-being of humankind.

More and more artificial systems are being deployed, affecting an increasing number of people. To ensure their well-being and to prevent these people from being harmed more than is necessary, it is imperative to develop systems that act in accordance with morals (i.e., to align them morally). In order to do so, we need systems with capacity for moral reasoning. Since machine ethics is the field of research that deals with equipping systems with that capacity, we must pursue this field. [11, 100, 346]

Motivation 1.1: Acceptance The deployment of many artificial systems is likely to be of benefit to us humans for various reasons, such as improving our standard of living [27, 100].

Autonomous Vehicle #2

It is often assumed that the widespread dissemination of autonomous vehicles (AVs) will be greatly beneficial: plausibly, they will reduce traffic jams, accidents, and pollution.

Consequently, we must ensure that such systems become accepted. However, for the systems to be accepted, we must be able to place justified trust in them [116], and, for that to happen, they must be trustworthy [264]. Giving systems the capacity to reason morally is one prerequisite for making them trustworthy and, thus, increasing their acceptance [23].

[4] A *pro-tanto* reason is a reason that, as it applies only to some limited extent, can be outweighed.

Motivation 1.2: Fairness Irrespective of whether systems are being equipped with capacity to reason morally, we as human beings have an obligation to design them in accordance with morals [11, 346]. Morality commands us, for instance, to make fair decisions. Artificial systems should, therefore, also make fair decisions. Nevertheless, it is often the case that, implicitly, wrong values are (accidentally) being programmed into a system, and the resulting systems are, therefore, biased. Systems based on ML only aggravate this problem.

Hiring System #1

In 2018 there was a media outcry because it came to light that Amazon had a hiring tool that discriminated against women [138]. The fact that systems based on ML tend to discriminate against women or minorities was not new at that point. Because ML systems learn patterns from historical data, the discrimination that is encoded in these data is also learned [52]. Furthermore, there is hardly any data that is free of discrimination. What was new about this scandal was that direct references to gender had been erased from the data. In other words, the system did not learn directly that gender equates to a low hiring chance, but that word choice, hobbies, and the like—as proxies for gender—correlate with low hiring chances.

Machine ethics can help to prevent implicit values from being programmed into a system, since implementing morals requires making values explicit [23, 24]. Notions of fairness are often about ensuring that protected attributes, such as race or gender, do not unduly influence an entity's decision-making process. For this reason, a system's potential capacity to reason morally must include the capacity to weigh the influence of protected attributes in a decision-making process. As a side effect of making the system fairer, the system also becomes more trustworthy and, thus, more accepted (here lies a link to Motivation 1.1).

Motivation 1.3: Benign Superintelligence The further we forecast the future, the more urgent it becomes to equip artificial systems with capacity for moral reasoning. This urgency is not only because the penetration of artificial systems into our everyday lives is unlikely to abate (it will most likely increase in pace), but also because more theoretical threats might arise, such as superintelligence: an entity that possesses intelligence far surpassing that of the brightest human minds. If a superintelligence emerges, then it is highly recommendable for this superintelligence to have capacity to reason morally [205, 346].

Superintelligence is sometimes referred to as "our final invention" (see, for instance, [53])—a label that has at least two possible readings: either the superintelligence will make human life more pleasant than it has ever been by finding solutions for all our problems, so that no further human inventions are necessary; or it will simply annihilate humanity. To increase the chances of the first scenario occurring, our best option is to equip artificial systems with capacity for moral reasoning.

Motivation 2: Moral Alignment The second family of reasons for researching machine ethics is concerned with the moral alignment of decision makers.

Ethics, as a research discipline, constantly progresses. Despite this progress, however, the discipline has revolved around roughly the same three families of normative theories for several hundred years. From a third-person perspective, ethics seems to be at an impasse: no family of normative theories is accepted across the board.[5] Many researchers see a possible way out of this dilemma in the emerging discipline of machine ethics. Engaging in machine ethics may improve the moral alignment of (machine or human) decision makers beyond current (human) standards [10, 11, 100]. We may become able to formulate better moral theories and, perhaps, even find the correct one (if there is such a thing).

Motivation 2.1: Improved Human Decisions Studying machine ethics can help to improve individual human decisions. Humans are prone to error. Not only do we exhibit many cognitive biases, but we also have limited cognitive abilities. Moreover, we often act out of self-interest and employ sloppy reasoning. [10, 23]

Machine reasoners are, in principle, free of such limitations. If designed carefully, a machine can be free of most human-like biases (however, see Motivation 1.2) [10, 23]. Additionally, the current computational power of machines is not only superior to that of humans, it is still rapidly improving [23]. Humans may use a machine to arrive at moral solutions just as they use pocket calculators to arrive at mathematical solutions [100, 469].

Digression #1

Although efficiency is the most prominent reason for the increasing use of ML, it is not the only one. In addition to being able to process immense amounts of data in a comparatively short time, ML was also hoped to eliminate problems of human decision making. Fatigued or hungry humans tend to err and make different decisions than their well-rested and sated counterparts. Machines, in contrast, can become neither hungry nor tired. Their decision making has an air of objectivity and faultlessness. That this conception is untenable, however, is nowadays widely accepted.

Motivation 2.2: Improved Human Morality The study of machine ethics can also help to improve human morality as a whole. Engaging in machine ethics can help us to formulate more consistent moral theories, and to reach consensus on moral dilemmas. [11, 23–26, 100]

In particular, algorithmizing normative theories can help to reveal inconsistencies in them [26, 377]. Given better knowledge of these inconsistencies, their respective theories can be amended (e.g., if the inconsistency is easily fixable) or abandoned (e.g., if the inconsistency demonstrates a fundamental problem of the theory).

[5] In 2014, a study found the following preference distribution of normative theories among philosophers: deontology 25.9%, consequentialism 23.6%, virtue ethics 18.2%, other 32.3% [80].

Furthermore, insights gleaned from studying differences between the kinds of entities engaging in moral deliberation (humans vs. machines) can also illuminate intriguing aspects of some theories and, possibly, contribute to the understanding of human morality as a whole [11] (we will discuss such aspects in Section 4.2). Let us illustrate these considerations.

Autonomous Vehicle #3

Philippa Foot's trolley cases [176] are among the most famous thought experiments in contemporary philosophy. With the advent of autonomous vehicles (AVs), these cases acquired even more fame: the originally hypothetical cases now come to reality in a slightly different facet. If faced with the decision to either knock down a group of pedestrians or deviate from its course and sacrifice its passenger, which option should an AV choose? To put it briefly, with the emergence of AVs, Foot's thought experiment became even more tangible than it already was. Finding an answer to the question of how the vehicles should behave is paramount for their successful deployment [76].

Sparrow, for instance, discusses in his "Turing Triage Test" the criteria that must be met to be considered a moral patient. He argues that applying traditional criteria, such as having self-consciousness or sentience, to artificial systems leads to counterintuitive results: Since superintelligences will, plausibly, fulfill these criteria to an even higher degree than humans, superintelligences will have a higher moral significance than humans. Moreover, since superintelligences will likely have almost unlimited possibilities for improvement, their moral significance could, eventually, outweigh that of humans to such an extent that the moral significance of humans becomes negligible. Since this is an unacceptable outcome, Sparrow argues, we need to rethink the criteria for attributing moral patiency.[6] [445]

Anderson, Anderson, and Armen give a concrete example where the algorithmization of normative theories can help to improve them. They consider William D. Ross' prima-facie duties and suggest how an artificial system might be able to resolve conflicts between competing duties (for more details, see Section 2.2.2 and Section 4.2.2). [24]

[6] Although it may serve as a good example, we consider Sparrow's argumentation to be significantly flawed. First, his argument seems to be close to an *argumentum ad consequentiam*. Rejecting a theory just because it has unacceptable consequences is a bad argumentative practice. It would be another story if the theory's implications were self-contradicting, but this is not the case here. Moreover (and second), the consequences outlined are very plausible and intended consequences of the traditional criteria for moral patiency. For example, it is sometimes assumed that we have more moral duties towards an intelligent great ape than towards a brain-damaged human [432]. Third, even if one denies that moral status comes in degrees, one might still argue that certain beings have more significant interests than others. Thus, for instance, one can argue that it is better to save the great ape instead of the brain-damaged person, not because the ape has higher moral status than the brain-damaged person, but because it has a more significant interest in having its life saved [79]. In this line of reasoning, the superintelligence would be more significant than a human, regardless of moral status. Fourth, we hold that the sketched consequences do not give us a reason to abandon the above criteria but rather a reason to prevent the emergence of a superintelligence. Finally (and fifth), the alternative criterion proposed by Sparrow is rather bizarre and difficult to apply. Furthermore, he does not really argue for it other than stating that the traditional criteria are inadequate. In summary, while the idea behind Sparrow's argument is laudable, its implementation is flawed.

To provide another example: Cave et al. suggest that artificial systems may be consulted for action guidance in situations where even human ethicists do not have a solution. [100]

Finally, Allen, Wallach, and Smit propose that machine ethics can help to stimulate new lines of inquiry in ethics. For example, considerations of what machines are capable of might lead to more profound reflections on what a correct normative theory is supposed to be. [11]

3.1.2. Reasons against Machine Ethics

We distinguish two families of reasons against pursuing research in machine ethics. In what follows (and as mentioned above), we will sometimes label these reasons as "risks" to demarcate them more clearly from reasons for pursuing this research discipline.

Risk 1: Insufficiently Moral Machines The first family of reasons against pursuing machine ethics research is concerned with its very possibility. More specifically, these reasons concern questions of whether machines can have sufficient morality for the task they are supposed to perform, indeed whether they can have morality at all.

Artificial systems (as they are understood here) differ fundamentally from humans. Even when we think of humans as a sophisticated type of machine, other contemporary machines lack essential qualities characteristic of humans [346]. Today's machines have neither sentience, nor (self-)consciousness, free will, or autonomy—all of which are essential human qualities that are commonly seen to qualify humans as beings with moral capacities [252]. Normative theories were, naturally, designed with humans in mind, or at least entities possessing some of the qualities mentioned above. Consequently, we need to investigate whether it is possible to apply results from ethics to machines, and whether it is possible to equip them with the capacity for moral reasoning. Some claim that this is not the case and that, consequently, the research discipline of machine ethics is misleading.

Risk 1.1: Lacking Moral Agency It is often claimed that machines cannot be moral. In contrast to humans, machines are not moral agents and, thus, cannot act morally. Machines not being moral agents is the case because machines lack the capacities mentioned in the previous paragraph. One simple formulation of this claim is that morality is about emotion, and machines cannot have emotion at all [10, 11, 346]. Arthur Schopenhauer, for example, famously claimed that compassion (arguably, a type of emotion) is the basis of morality [418]. Similar arguments can be brought forward for the other capacities (free will, autonomy, etc.).

Even if morality does not presuppose emotion (or some other capacity that machines lack), it is plausible that the presence of emotions (or said capacity) is at least one factor contributing to making praiseworthy moral decisions [469]. For this reason, it is questionable whether an acceptable level of morality can be achieved without specific capacities that machines lack. Consequently, it does not make sense to pursue research into machine ethics.

Risk 1.2: Unacceptable Level of Morality Let us continue with the line of reasoning set out in the preceding risk. We have to deliberate whether machines can achieve an acceptable level of morality at all. For machines to be accepted, it is plausible that a higher level of morality is required than that for humans [10, 11, 100]. This level of morality might very well be unattainable. In the literature, we can identify at least two considerations that support the view that machines require a higher level of moral capacities than humans (see [100]).

First, systems have high scalability. Artificial systems can easily be deployed on a large scale. For this reason, the algorithms employed in these systems effect far-reaching consequences. For example, the algorithms employed in Google, Facebook, or Amazon affect billions of people worldwide every day. Even on a smaller scale, the decision to use a specific system often leads to many people being affected by it [359]. In contrast, it is not easy for humans to come into a position where they can exert a similarly far-reaching influence. [100]

Second, systems have low predictability. Systems can fail in ways in which no human would have failed. For this reason, they can also come to decisions that are morally reprehensible in ways that no one could have foreseen. A human-like moral capacity might, therefore, not be sufficient for machines. Since it is plausible that a sufficient level of morality in machines is unattainable, we should not pursue machine ethics research. [100]

Risk 1.3: Computational Limitations Finally, it may be impossible to implement an adequate moral theory in a machine due to computational limitations [11, 88, 100].

Let us take some common normative theories as examples. Some versions of consequentialism require the decision maker to include all living beings in the universe in the deliberation process. Fulfilling this requirement seems to be computationally infeasible: calculating the effect of an action on each of these entities in real time seems impossible [10, 11].

Similarly, some researchers argue that it is hard to impossible to calculate the type of abstraction necessary for considerations required by the categorical imperative (e.g., [10, 377]).

Combined with the idea that machines need a level of morality higher than that of humans, this reason for not pursuing the research discipline of machine ethics becomes even stronger.

Risk 2: Detrimental Consequences The second family of reasons against pursuing machine ethics research is concerned with the detrimental consequences that this pursuit could bring about. More specifically, engaging in machine ethics may aggravate the matter and lead to the opposite of what it is intended to bring about.

Although, at first glance, it seems to be a good idea to equip machines with the capacity to reason morally, doing so may also lead to consequences that are not desirable. There are at least two facets to how this may come about. First, it may be that equipping machines with said capacities requires other capacities of machines, the integration of which could result in unintended and undesirable consequences (e.g., increased corruptibility). Accordingly,

trade-off needs to be carefully evaluated to determine whether the whole endeavor of machine ethics is still worthwhile (Risk 2.1.1 and Risk 2.1.2). Second, equipping machines with the ability to reason morally may, in itself, lead to harmful consequences (Risks 2.2.1–2.2.4).

Risk 2.1.1: Increased Corruptibility Moral capacities in a machine require an explicit representation of moral considerations (e.g., in the form of moral principles). Although this requirement is well-intentioned, it may also make the machine more corruptible [100, 474].

There are several ways in which this may come to be. Take hackers as an example, who may corrupt the moral faculties of a machine, making the system immoral. The problems can, however, begin even earlier. Moral faculties may also be used to make a system deliberately as evil as possible.

Somewhere between these two extremes lies another problem. The corruption of a machine can happen unintentionally, through coding errors. Ensuring that complex systems are error-free is difficult, if not impossible. So, if equipping systems with capacities to reason morally leads to such problems, we should not do so.

Risk 2.1.2: Moral Patiency If endowing systems with moral capacities leads too far, these systems can become moral patients [100]. If, for instance, we equip machines with emotions in order for them to be able to reason morally, we may get systems that have a moral status. We must not use systems with moral status in the same way that we use systems currently, because systems with the status of moral patients have certain rights that we must not violate [197, 224]. Furthermore, we must also take their interests seriously.

As our modern society depends on using artificial systems without considerations of their interests and rights (because, as of yet, they likely do not have any), a change in this status would significantly reduce their utility for humans. Consequently, we have an interest in not designing artificial systems that may classify as moral patients. If equipping systems with capacities for moral reasoning requires capacities that lead to such problems, we should not do so.

Risk 2.2.1: Bad Moral Performance Even with moral capacities, machines may yet exhibit worse moral performance than humans. There is a non-trivial risk that machines come to morally unacceptable conclusions that humans would have foreseen [11]. Such conclusions can arise, for example, due to the system working with false premises [88, 100]. The system's sensors might fail, delivering incorrect information to the system's decision-making processes.

Another way systems may fail is when their decision-making procedures do not lead to any action or arrive at incompatible actions, in cases where humans would likely have reached an acceptable conclusion [11]. In general, humans may identify incompatible courses of action more easily than machines, and they reliably manage to arrive at a course of action [11, 100].

In addition, equipping a system with moral capabilities is likely to result in overall poorer performance because the system has to perform more processing steps to come to a result (e.g., when the moral capabilities are implemented as additional constraints in the system's decision making) [11]. So, if machines equipped with capacities for moral reasoning exhibit worse moral performance than humans, then, at the least, we should not be too zealous about equipping them with such capacities.

Risk 2.2.2: Responsibility Gap As mentioned earlier, system deployment sometimes goes hand-in-hand with problems of attributing accountability or responsibility [11, 63, 100, 323].

Autonomous Vehicle #4

In the case of AVs, several parties are potentially responsible for an accident. First, the person sitting in what was formerly the driver's seat. Are they responsible because they did not intervene to prevent the accident? Depending on the AV, intervening may not even be possible. Furthermore, only a minority of future accidents may be preventable by human intervention. Is the car's vendor responsible because they sold the car? Is it the watchdog organization (e.g., TÜV in Germany) that certified the car? There are many other candidates: a programmer (and which programmer, of the many involved?), the manufacturing company (which people in it?), etc. While some candidates are less plausible (e.g., the vendor), there remain many to whom it is reasonable to attribute responsibility.

In modern bureaucracies, established procedures often distribute responsibility so widely that no one person can be identified to blame for an accident [245]. Deploying machines with moral capacities could make the attribution of responsibility even more diffuse than it already is, for at least three reasons.

First, the increased complexity of the decision-making processes may make it more challenging to determine whether the accident was caused by a flaw in the system's code or by something else (e.g., a sensor failure). Furthermore, determining whether the moral deliberation process on which the decision was based is flawed faces similar problems.

Second, the design and implementation of moral faculties introduces more parties to potentially shift blame to (e.g., the policy-makers or ethicists who designed them) and could invite other parties to do a sloppier job [100].

Third, the presence of a supposedly moral machine during an accident may invite one to shift the blame to this machine [79, 132, 423]. It is easier to use the machine as a scapegoat than to take responsibility as a human.

Taken together, if pursuing the research discipline of machine ethics makes the attribution of responsibility more difficult, then we should not be especially zealous in doing so.

Risk 2.2.3: Value Imperialism The scalability of artificial systems can, plausibly, lead to some kind of value imperialism [100]. When systems are endowed with specific moral capacities, these capacities are, to some extent, fixed. Different cultures, however, endorse different moral values (see Autonomous Vehicle #5).

Autonomous Vehicle #5

In a recent study [46], researchers presented several dilemma situations that an AV had to face, to people all over the world. These situations differed, for instance, with respect to the age, gender, number, or social status of the persons who could be spared or sacrificed. The study's findings revealed remarkable differences between different cultures in terms of the preferences they exercised. Participants from Western and Southern countries, for instance, preferred to spare younger people, a preference that was utterly absent among participants from Eastern countries.[a] Furthermore, participants from Southern countries preferred to spare females, a preference that was not very pronounced among participants from Western and Eastern countries.

[a]Western: e.g., Germany, USA; Southern: e.g., Chile, Algeria; Eastern: e.g., Japan, China.

Against this background, equipping systems with a particular capacity to reason morally can lead to violations of cultural identity. A system that acts according to a consequentialist picture of morality (and, for instance, usually spares younger people) may be the obvious choice for certain countries, but it could be considered immoral in other countries.

This problem becomes even more apparent when we look at it not only from a contemporary perspective but extrapolate it to the future. Values and moral notions change over time. The ancient Greeks considered slavery as allowed; today, it is commonly regarded as impermissible. If the ancient Greeks were to embed their values in a computer program still in use today, such a program would likely be deemed immoral.

Although the problem of values changing over time may initially sound rather unimportant, it is a problem often discussed regarding superintelligences. We have strong reasons to make a superintelligence act in accordance with morals (as outlined in Motivation 1.3). By doing so, the superintelligence may have human-like morality for a few years, but not forever. It is plausible that the superintelligence's picture of morality and ours will, at some point, diverge, just like the picture of morality the ancient Greeks had is different from ours. For these reasons, one should refrain from equipping systems with capacities for moral deliberation if this leads to value dominance (regardless of which picture of morality is the better one).

Risk 2.2.4: Undermined Human Agency Equipping machines with moral capacities may undermine human agency [100]. Machines with such capacities may support human incompetence by correcting their (moral) mistakes. In other words, the fear is that our human moral skills will erode because we are not required to hone them. [212]

Just like our capacities for mental arithmetic are likely to deteriorate through dependence on pocket calculators, our moral judgment skills may deteriorate through dependence on machines to support our (moral) decision making. In the worst case, this development could lead to people no longer developing (sufficient) moral skills. If we start to rely on machines for moral judgments or our judgments have to be compared to those of machines, there is no motivation to (further) develop our abilities.

Overall, the fear is that, in the end, humans would degrade to being the machines' compliant vicarious agents and, thus, forfeit significant parts of their agency. As a result, humans may not be able to remain responsible decision makers because they lack the capacities to make informed and well-founded decisions (here lies a link to Risk 2.2.2).

3.1.3. Refuting the Reasons against Machine Ethics

At this point, we have quantitatively listed more reasons that speak against pursuing the research discipline of machine ethics than in favor of it. Qualitatively, however, the ratio is reversed. Most of the reasons against engaging in machine ethics seem strong at first, but can be easily toned down, if not entirely refuted. While doing this, we will also start to sketch, more thoroughly, our picture of what machine ethics should look like.

Risk 1: Insufficiently Moral Machines The general rationale behind refuting this family of risks is that, while it is true that modern machines lack the capacities needed for being moral in the strong sense humans are, we advocate that this is not problematic. On the contrary, this is even desired. Indeed, machine ethics would be misguided if it solely sought to make machines moral in this strong sense. Many of the problems described in the last section arise only when one tries to do just that. However, we propose that a more modest goal (viz., moral alignment) can also do the job.

Machine ethics can be seen as an answer to the ever-increasing influence of artificial systems on human life. To make this influence as beneficial as possible, we argue that we do not need machines to be moral in the same way humans are. Equipping machines with an explicit representation of moral considerations and a decision-making procedure that takes these considerations into account should be sufficient to raise the behavior of these machines to an appropriate level.[7]

Risk 1.1: Lacking Moral Agency As just described, morality in the strong sense might be unattainable for machines (more than that, it is even undesirable). Regardless of what capacities are prerequisite to having morality on the level of agents, it does not matter whether a system has them. In our eyes, the goal has never been to create artificial systems with some sort of personal morality—as opposed to designing them to perform morally desirable actions.

[7] Admittedly, this kind of argument seems, at first glance, to presuppose a consequentialist picture of morality. However, the argument is still sound if one reads "beneficial" with a reasonable picture of morality in mind. Take virtue ethics as an example. It is plausible that artificial systems acting according to morals help us humans to live a good life and develop ourselves. Similar arguments can be made for other moral theories.

When one tries to build an ethical component into a robot, the goal is not to create a morally "good" robot; rather, the goal is to lay the foundation for machine behavior that, in the best case, would regularly be classified as morally excellent if it came from a human.

As mentioned, machine ethics should pursue a high moral alignment of artificial systems. To push the behavior of machines at least a little bit in a direction more advantageous for humans is already sufficient to justify machine ethics research.

Risk 1.2: Unacceptable Level of Morality If one accepts that the goals of machine ethics should be lower than is often proclaimed, one can easily see that this risk is, in fact, groundless. It may be the case that some systems will (or even should) not be accepted if their morality does not surpass human levels, but this poses no real problems for at least three reasons.

First, it is arguably the case that these systems should not be deployed at all. Take the services offered by Facebook, Google, and Amazon as examples. Many claim that their predominance in certain areas is not desirable. Similarly, it may be that artificial systems, in general, should either not be used on a large scale or at least be subject to thorough scrutiny before this happens. In this line of thought, machine ethics can help us to make systems more moral, but that should not preclude them from being scrutinized by regulators. A system should only be used if it meets certain standards. Machine ethics can help achieve these standards, but it does not have to be the only ingredient to do so.

Second, machine ethics is not intended to mitigate the fallibility of machines completely. Again, machine ethics can be an essential ingredient in preventing systems from failing. Alas, it is difficult to guarantee that machines are faultless, and it is also the case that moral capacities may prove to be a new source of potential failure. Nevertheless, it is plausible that, overall, morally aligned systems, if carefully designed, will exhibit significantly better (moral) behavior than those without moral alignment.

Finally (and third), artificial systems will continue to increasingly infiltrate our daily lives, whether or not they have moral capacities. Against this background, it is better to make them as moral as possible than not moral at all.

Medical-Care Robot #2

One of the biggest problems facing many of today's societies is demographic transition. As populations age, many countries are experiencing severe problems with eldercare. Among these problems is the scarcity of caregivers. To address this problem, robots are increasingly being used to perform caregiving tasks. For many people, there is no alternative, as the nursing professions are chronically understaffed and more needed than ever. For this reason, it is arguably better to employ robots to do this work than not to have it done at all. With this in mind, these robots should have at least some capacity for moral reasoning, as their field of application comprises direct contact with humans.

Risk 1.3: Computational Limitations Machines may be computationally limited, but humans are even more so [286]. Therefore, it seems to be a bad reason to stop pursuing machine ethics because of computational limitations. Indeed, if one accepts that machine ethics should merely aim to increase the moral alignment of artificial systems, it makes no sense to consider approaches to morality that defy implementation in a machine (e.g., because they are computationally too complex). The upshot here is that there is no need to implement any of the established normative theories. Of course, doing so seems to be a straightforward (and deceptively easy) way to achieve moral alignment. However, it is also subject to specific problems, some of which we will discuss later (in Section 4.1.1).

For this reason, it may be even more advantageous to develop theories that are specifically tailored to machines. A computationally too demanding theory for a machine is probably also too demanding for a human being. For example, the computational limitations of humans are often used to criticize consequentialist theories (see Section 4.2.1). An artificial system is more likely to be able to perform the required calculations, both more quickly and more accurately than a human [23]. After all, machines are becoming more powerful by the day, and today's computational limitations (e.g., storage) are likely to be alleviated in the future.

Risk 2: Detrimental Consequences If we set the goal of machine ethics as moral alignment, both types of reasons we find in this family can be refuted. First, many problematic additional capacities (e.g., sentience) are not required if we adopt the view of machine ethics that we are defending here. Thus, risks of the first type can be averted. Second, by adopting the view described, most of the negative consequences listed can be avoided. Thus, risks of the second type fail.

Risk 2.1.1: Increased Corruptibility Regardless of whether artificial systems have moral capacities or not, programming mistakes, malicious design, or hacking can always happen. The possibility of such occurrences is not a decisive argument against pursuing machine ethics research. Research in computer science has developed tools and ways to cope with such problems. For example, research in verification is concerned with ensuring that a program only exhibits the intended behavior [50, 298]. Furthermore, research in software engineering aims to minimize the error rates in systems [504]. A final example is the rapidly growing research in cybersecurity, which focuses on preventing malicious attacks on systems [268].

These are just a few examples of research trying to prevent systems from being maliciously exploited. As we will discuss later (in Section 3.2.1), research in machine explainability promises to be a helpful addendum to machine ethics when it comes to this reason against it (see also Section 9.2.1). If the moral capacities of a system are manipulated, it stands to reason that such a manipulation is often detectable with the help of machine explainability, plausibly even more effortless than other ones.

Risk 2.1.2: Moral Patiency If moral alignment is accepted as the goal of machine ethics, the danger of artificial systems becoming moral patients is mostly mitigated. Morally aligning systems does not require substantially different capacities from those that machines without such an alignment have. Consequently, there is little benefit in machines having capacities that could qualify them as moral patients. To sum it up, if moral alignment does not require new capacities, the question of moral status is detached from making systems moral.

Admittedly, there is a possibility that, at some point, there will be artificial systems that qualify as moral patients. Indeed, there is concern that if the speed of progress in AI remains as it is, advances could inevitably lead to some systems being classified as moral patients. However, this is more of a reason to oppose AI research than machine ethics research. As a matter of fact, some researchers who fear the emergence of an artificial general intelligence (that could evolve into a superintelligence) are already calling for a (temporary) moratorium on specific kinds of AI research [460].

Risk 2.2.1: Bad Moral Performance To reject a technology just because it might fail in situations where the old technology did not fail is a threat to beneficial scientific progress. Numerous inventions (many of which are widespread) come with this putative drawback.

Autonomous Vehicle #6

As previously mentioned (in Motivation 1.1), the widespread dissemination of autonomous vehicles will likely bring about tremendous advantages. Nevertheless, autonomous vehicles will be involved in different types of accidents than conventional vehicles, and, consequently, different types of casualties will ensue.

However, this does not mean that efforts in developing and designing autonomous vehicles have decreased. Instead, it has led to the formation of ethics committees to get the best out of this newly emerging technology while avoiding the worst.

Let us look at another example that might be easier to identify with.

Autonomous Vehicle #7

Although there are, arguably, instances when people would not have died if it were not for the safety belt, we are required to wear a safety belt while driving in a car. The general benefits of wearing a safety belt outweigh the costs involved.

Both vehicle autonomy and safety belts are likely to improve the situation for many people without placing specific groups of people at a disadvantage. Similarly, we should not reject progress in machine ethics just because some cases will be worse than before. In the end, the only thing that counts is that overall, our world will be a better place.

Risk 2.2.2: Responsibility Gap The problems arising from a diffusion of accountability and responsibility are, admittedly, hard to resolve at this point. We will return to these problems in Section 3.2.1 (and, in particular, in Section 9.3) while discussing how machine explainability contributes to machine ethics. Nevertheless, we advocate that machine ethics does not make the case worse than before, and we do so for two reasons.

First, falsely designed moral capacities can potentially be exposed in relevant cases. In cases of accidents, it is sometimes possible to reconstruct the past decision-making processes of the system. Therefore, it may be possible to identify whether the system's implementation or its moral capacities caused the failure. If this is true, the allocation of responsibility is, at least, not worse than it already was.

Second, some failure cases can be prevented by machine ethics. Recall Motivation 1.2: machine ethics can help to prevent the embedding of implicit value in systems. Such implicit embeddings contribute to a large class of failures where the attribution of responsibility is particularly complicated. Thus, containing this class of failures is likely to make the overall allocation of responsibility easier.

Risk 2.2.3: Value Imperialism Before describing a simple way to escape value imperialism, let us first reflect on the relativity of morals. For value imperialism to be a bad thing, morals would have to be relative. Although different cultural groups in the world seem to favor different values (see above), whether morals are really relative is highly controversial [49, 195]. The more common picture is that one normative theory is objectively correct [25, 80] (the controversy in this case, however, is about which one is correct).

Nevertheless, even if morals were relative, this does not pose a problem to pursuing the research discipline of machine ethics. For example, one could factor in the country (or cultural region) a system is deployed in when designing its moral capacities. Even if the system is supposed to be deployed in several countries, different moral capacities may be programmed into it, changing according to the country or region in which the system is currently located.

Moreover, some kind of value imperialism is already present in systems without explicit moral reasoning components since it is likely that the programmers' biases are implicitly embedded. Machine ethics can make it possible to prevent such implicit biases to a certain degree, as previously argued (see Motivation 1.2).

Furthermore, the ability to adapt the moral faculties implemented in a system can also make it possible to evade problems with superintelligences. First, the ability to adapt its own moral faculties may prevent the superintelligence from rigidly adhering to its initial picture of morality. Second, when a divergence of moral beliefs between humans and superintelligence is identified, there is a starting point for closing such a gap.

Risk 2.2.4: Undermining Human Agency This risk is also much easier to dismiss if one takes explainability into account (see Section 3.2.1). Nevertheless, pursuing the conception of machine ethics that we defend already escapes this risk to some degree. Since machines are

in no way intended to replace us humans in decision making, the likelihood that human skills will erode decreases. Furthermore, perhaps we should even regulate machines in a way that precludes them from being able to replace humans in certain areas.

3.2. The Advantages of Machine Explainability

Most, if not all, reasons against pursuing machine ethics research can be refuted quite easily. Given the potential benefits that this research may yield, it is well worth pursuing. We will see so especially in this subsection, where we will discuss how machine explainability can further augment machine ethics.

Results from machine explainability can help in three ways. First, they can be used to amend some of the disadvantages of machine ethics. Second, they can be used to augment some of the advantages. Finally, they can even engender entirely new advantages. Consequently, if we have reason to research machine ethics, we also have reason to research machine explainability as well.

Machine explainability will be the focus in the third part of this thesis; this chapter only aims at illustrating the advantages of machine explainability in connection with machine ethics. In a nutshell, it can be stated that machine explainability is concerned with making various aspects of an artificial system understandable to a stakeholder (see also Section 1.2). Whether it is the visible behavior, the algorithm on which this behavior is based, or the input needed to produce a certain behavior, producing explanations of it is a legitimate goal pursued by research in machine explainability.

3.2.1. Amending the Disadvantages of Machine Ethics

Machine explainability can help mitigate the risks mentioned before. This holds for some risks more than for others. We will focus on the risks that can be best mitigated in what follows. In particular, these are the risks that we did not (fully) mitigate previously.

Risk 2.1.1: Increased Corruptibility The more complex artificial systems are, the more opaque their reasoning processes become. Take deep neural networks (DNNs) as an example. Systems that employ DNNs for their decision-making procedure can fail in ways that no human would have predicted (see [192] for some compelling examples). Furthermore, their sub-symbolic processing makes it hard for humans to understand their inner workings and, thus, hard to fix them.

One of the motivations for research in machine explainability is to make systems better debuggable and maintainable [73, 342] (see also Section 9.2.1 and Table 12). Explanations of a system's inner workings can help pinpoint sources of failure and, thus, enable developers to fix them. So, even if machine ethics leads to new sources of failure, machine explainability has the potential to aid us in finding and eliminating them.

Risk 2.2.1: Bad Moral Performance In the last section, we noted that machines, even when endowed with moral capabilities, might perform worse than humans. However, we argued that this is tenable, as the goal of machine ethics should be to improve the moral alignment of machines as much as is possible, and not necessarily to perfect it. In this line of reasoning, the performance gap between humans and systems, while unfortunate, is acceptable in many cases.

Naturally, this does not mean that we should not try to bridge the gap. This is where machine explainability comes into play. Where unacceptable outcomes occur, explanations can help identify where a machine's moral capacities are defective and need to be adjusted. Machine explainability allows us to continuously improve the system and its moral faculties.

Additionally, even in cases where the errors are based on false premises, such as sensor failures, and improvements in the moral faculties are to no avail, machine explainability may help to identify possible ways of failure and, based on these, help the development of safeguarding strategies (e.g., safety modes when sensor data deviates from specific standards).

Risk 2.2.2: Responsibility Gap One of the central motivations for pursuing machine explainability is to be better able to attribute responsibility [342, 365] (see also Section 9.2.1 and Table 25). Indeed, this topic is of such great importance that a whole subsection in this thesis will be dedicated to the connection between machine explainability and responsibility (viz., Section 9.3). At this point, we will, thus, only mention essential thoughts.

As already outlined, the widespread introduction of artificial systems brought about the so-called "responsibility gap" [323]: where it used to be clear who was responsible for an accident, with artificial systems in play, this responsibility diffuses to several parties in such a way that, often, no one can justifiedly be held responsible. Especially in scenarios where much is at stake (e.g., human lives), such conditions are undesirable for many reasons (e.g., justice).

Even without factoring in new parties (to which one could potentially attribute responsibility) that machine ethics may bring into the mix, there are already other philosophical problems associated with problems of responsibility attribution. One fundamental problem that seems to cause many other problems is that humans lose their ability to make competent decisions. When required to act upon the outputs of a system whose internal processes they do not understand, humans degrade to compliant enforcers. They cannot justifiedly decide for or against a system's output. This circumstance is one of the reasons that makes it so hard to justifiedly attribute responsibility to them: they could not have done better.

At least in this case, machine explainability can help. The idea is that humans are put back into positions where they justifiedly bear responsibility. Human-in-the-loop scenarios, where a human makes the final decision based on an artificial system's recommendations, are well-suited for this purpose. When the system can explain its recommendation, the human operator can reconcile this explanation with their view and competently decide. Thus, the

operator can justifiedly reject or accept the system's recommendation. For this reason, the operator can bear the responsibility.

Machine explainability can also help in other cases. As seen in the last risks, machine explainability can help to pinpoint where a fault originates. That being done, it is then possible to better pinpoint who is responsible. If an explanation reveals a programming mistake, the developer is most likely responsible. Similarly, if an explanation reveals a flaw in the moral capacities of the system—a flaw that does not originate from a programming mistake—then the ethicist designing these capacities is most likely responsible.

Risk 2.2.4: Undermined Human Agency Making machines explainable can help preserve human agency. We have already outlined how explanations can help humans remain responsible in scenarios where they act upon a machine's output (and we will explore this topic further in Section 9.3). In short, explanations let humans (regain) control over a situation. As a consequence, humans remain responsible in this situation and uphold their agency.

However, the positive impact of machine explainability can go even further and may prevent a potential skill erosion. By providing explanations about the moral considerations involved in a decision-making process, the system can educate humans [222, 339] (see also Section 9.2.1 and Table 13). Perhaps, at some point in time, machines will acquire a level of morality above that of humans. In such a scenario, it would be plausible for humans to gain moral insights by receiving explanations of the moral considerations that play a role in various scenarios. Even in a more pessimistic scenario, where skills erode because humans overly rely on machines taking over their (moral) decisions, explanations may at least decelerate such an erosion.

3.2.2. Augmenting the Advantages of Machine Ethics

Machine explainability makes it possible to significantly augment each of the reasons for pursuing the research discipline of machine ethics.

Motivation 1.1: Acceptance Besides the attribution of responsibility, acceptance of artificial systems is another central motivator for machine explainability [73, 173] (see also Section 9.2.1 and Table 6). Machine explainability can promote the acceptance of artificial systems in at least two ways.

First, it is often argued that humans are more likely to accept entities that can provide explanations and justify their behavior. If a person can see that the output of an artificial system is based on a valid reasoning process, they are more likely to accept such a system [70, 116, 280]. Based on this argument, one could claim that an explainable system with a machine ethics component is more likely to become accepted than a system without one.

Although the connection between explainability and acceptance is often drawn, things are not that easy. Explanations can contribute to acceptance, but they do not have to. For example, if the ethical component of a system depends on moral principles that one does

not share or that one even condemns, then an explanation that exposes this fact will reduce rather than increase acceptance. A similar relationship can be observed with regard to trust, which we will touch upon later (in Section 9.2.1). At this point, it should be noted that, even if acceptance is not increased in some cases, this is not a disadvantage: what is ideally achieved by explanations is justified and well-calibrated acceptance, as opposed to blind acceptance.

Second, machine explainability can help to shed light on accidents. If a system is equipped with capacity to explain its reasoning process, this can help in case of accidents. Not only does it make the attribution of responsibility easier (as previously argued), but it also helps to maintain people's acceptance. Systems that fail for unknown reasons are barely worth our acceptance. Traceable failures, on the other hand, should not significantly affect our acceptance.

Motivation 1.2: Fairness Fairness is another central motivator for machine explainability [1, 4, 54] (see also Section 9.2.1 and Table 16). While machine ethics requires us to make implicit values explicit, machine explainability goes one step further. Notions of fairness are often about ensuring that protected attributes such as race or gender do not unduly influence an entity's decision-making process. As machine explainability aims to uncover the precise workings of a system's reasoning process, such influences are, ideally, also uncovered.

At this point, one can see the synergy between machine ethics and machine explainability particularly well. Where machine ethics may require implicit values to be laid open, machine explainability helps uncover how these values were used to make decisions.

Nevertheless, machine explainability does not only aim at cases where the protected attributes are laid open. In DNNs, for instance, the influence of many values is not easy to understand. Furthermore, some can be proxies for protected attributes. In other words, even if an attribute like gender is denied to the algorithm during its training phase, other attributes, like word choice or hobbies, can correlate with gender (see Hiring System #1). Tracing the influence of different values through explanations can help to reveal such implicit influences. Based on the insights gained, systems can be improved and, thus, made fairer.

Motivation 1.3: Benign Superintelligence As the term "superintelligence" implies, a superintelligent entity far surpasses human intelligence. For this reason, the doings and thought processes of a superintelligence may become incomprehensible and elusive to humans. Consequently, it may not be evident whether the doings of a superintelligent entity are to our benefit or not. The larger implications of its actions may simply not be apparent to us.

If, however, a superintelligence is endowed with the ability to explain its doings in human-comprehensible terms, such an analysis may become possible. In general, there are many advantages associated with a superintelligence that can explain its doings and thought processes. For instance, it can convey scientific knowledge to humans and, thus, educate them.

Motivation 2.1: Improved Human Decisions When an artificial system is used to assist a human in making moral decisions, this can also help to improve the human's morality. To do so, however, the human has to have access to the salient considerations that the machine used to come to its recommendation. If the system is able to give explanations, such access is provided. As previously argued, the system can educate humans by thoroughly explaining why the given course of action is seen as morally good.

Motivation 2.2: Improved Human Morality Education achieved by explaining how a particular recommendation or course of action came about does not have to stop on an individual human level. Artificial systems will face situations in which there is a moral dilemma, and they must make choices in these situations. By explaining their solutions, we might learn something about morality as a whole and rethink our approach to similar situations. This can go so far that (new) standards may even be established.

3.2.3. New Advantages for Machine Ethics

Machine explainability does not only avert risks of machine ethics and augments its advantages, but is also beneficial on its own. Indeed, machine explainability promises to bring about many advantages, some of which we will discuss in detail in Section 9.2.1. In this section, we want to elaborate some advantages that are specific to machine ethics.

Advantage 1.1: Machine Ethics Acceptance In addition to the acceptance of *systems* based on the human psychological need for explanations (see augmentation of Motivation 1.1), machine explainability can promote the acceptance of machine ethics *itself*.

Medical-Care Robot #3

Imagine a medical care robot in a hospital's intensive care unit. Although it witnesses a critical scene in which a patient is about to die, the robot does not help (despite having the ability to help). As onlookers, we are aware that the robot is endowed with capacities for moral reasoning. For this reason, we wonder why the robot acted the way it did. Furthermore, our credence in the robot's moral capacities is lowered.

Now imagine that the robot could explain itself and state that it was well aware of the critical condition but estimated that its battery charge would not have been sufficient to help the patient. For this reason, it turned around and called the doctor. So, if we knew why the robot acted the way it did (and we thought it made sense to do so), we should be more accepting of it and stop questioning its moral capacities.

In many cases, the sole assurance that a system acts in accordance with morals may not suffice. In cases where a system's behavior seems immoral from an outside perspective, it is crucial to be able to distinguish malfunction from proper behavior. It is even more

important to get to the bottom of this behavior, to understand how it came about. Machine explainability can help with both. By understanding how the system came to its behavior, we can distinguish intended from unintended behavior. In cases of intended behavior, we may judge it as reasonable and, thus, restore our trust in the system. In cases of malfunction, we may become able to fix the faults.

Advantage 1.2: Improved Machine Morality As we have already depicted, machine explainability can help us to improve systems. This improvement is not limited to fixing faults in their programming or enhancing their main functionality: as the capacities for moral reasoning are part of the systems, machine explainability also enables the capacities' improvement. By obtaining explanations of what led a system to exhibit specific behavior, be it an acceptable or unacceptable one, we can better analyze whether the capacities are well-designed or not. In this way, it may be possible to close the moral gap between humans and machines and, hopefully, produce machines that are (in their behavior) even morally superior to humans.

Advantage 1.3: Enriched Machine Ethics Machine explainability can be seen as a part of machine ethics itself [100, 483]. As we have demonstrated, machine explainability can help machine ethics in various ways. Some of these ways are genuinely important for machine ethics. For instance, ascribing responsibility and upholding human agency are genuine moral concerns (for more examples, see Section 9.2.1 and Appendix C). For this reason, machine explainability is not simply a convenient addendum to machine ethics, but an integral part of it. With this in mind, let us come back to our vision of machine ethics. Equipping machines with the ability to explain themselves, or devising methods to do so from the outside, is pivotal for pushing them in the direction of being more moral.

Our discussion shows how machine ethics and machine explainability are deeply intertwined. On the one hand, machine ethics needs machine explainability in order to reach its full potential. On the other hand, machine explainability can profit from machine ethics as machine ethics offers points for machine explainability to hook in. In order to more fully illuminate this relationship, we must first finish outlining our conception of machine ethics by considering its implementation in the next section.

4. Implementing Machine Morality

In the last section, we implicitly argued for our view on machine ethics by pointing out how adopting it can avoid many problems associated with machine ethics. To summarize, we think that, for the time being, machine ethics should not aim directly at *truly moral* decision making. Rather, the most pressing task of machine ethics is, currently, to align systems morally by finding an acceptable *morally constrained* means of decision making.

Having this concept of machine morality in mind, the question is how to implement it. The implementation must allow for clear formal guarantees that restrict the behavior of an autonomous system in a way that makes the system significantly morally better [146], and still allows it to continue to function as intended. The goal is, thus, an overall morally desirable system that remains useful.

In this section, we will take a closer look at how to realize this goal. To be precise, we will present our approach more thoroughly and provide more arguments in support of it. In doing so, we will also illustrate some of the advantages of machine ethics in more detail.

4.1. A Principle-Guided Approach to Implementing Morals

As illustrated in Section 2.2.2, many approaches in machine ethics are profoundly inspired by traditional moral theories, and some even attempt to implement one of them directly. In this section, we will argue that there is no actual need to do so; in fact, the opposite is the case. This argument provides support to our view that moral alignment should be the goal of machine ethics. Subsequently, we will argue that a *principle-guided* approach is a good choice for such a conception of machine ethics.

Moral alignment does not presuppose that machines qualify as moral agents since they do not need to have the capacities for genuine moral decision making. As we have argued, it is not desirable for them to have these capacities. It does need to be mentioned, however, that moral alignment does not guarantee perfectly moral behavior. Admittedly, in some cases, moral alignment might fail to achieve desirable standards altogether. With these two thoughts in mind, we will call our view a "reduced concept of machine morality" to highlight the constrained domain over which it applies: this approach is not fool-proof, and is not intended to be.

4.1.1. Reasons for a Reduced Concept of Machine Morality

In order to argue for our reduced concept, we argue against a full-fledged concept. In other words, we argue against the view that machine ethics is about straightforwardly implementing a traditional normative theory. There are several compelling arguments for our view.

Normative Plurality There is no agreement on the correct normative theory or whether such a theory exists at all. People have argued about the first question for more than 2,000 years, and we cannot seriously hope for a consensus within the next few years.

As described in Section 2.1.2, there are currently three major families of normative theories: consequentialist, deontological, and virtue. All of them come with their own advantages and problems. Philippa Foot, for example, prominently emphasized the tension between consequentialism and common sense (in [176]).[8]

We will discuss each of these families later (in Section 4.2.1–Section 4.2.3) and analyze whether they are fit to be implemented into an artificial system. As of yet, however, no family of theories has produced a member that exhibits a distinct advantage over the others.

Missing Moral Agency Conventional normative theories have emerged with humans in mind. Although this seems to be, at first glance, no decisive objection to implementing them in an artificial system, it still raises several non-trivial questions. The standard view is that only moral agents can exhibit acts to which we may predicate terms of morality [100]. In order to praise or blame some entity for its behavior, that entity must have certain qualities.

Sentience, (self-)consciousness, free will, or autonomy are commonly seen as prerequisites to being genuinely considered a moral agent [252] (see also Section 2.2.1). As of yet, however, there are no artificial systems that possess even one of these qualities. For this reason, it is questionable whether implementing a conventional normative theory is the best approach to align artificial systems morally. Indeed, it might be more beneficial (as we defend) to devise new moral theories explicitly for entities that do not have moral status.

Self-Refutation Implementing traditional normative theories may be undesirable from the point of view of these very theories. This undesirability may be due to the fact that such an implementation runs counter to what the theory is trying to accomplish.

Take utilitarianism (as important branch of consequentialist theories) as an example. David Hodgson argued that a society solely consisting of perfect utilitarianists would be, in fact, not desirable from the viewpoint of this theory [228] (discussions of his argument can be found in [315] and [431]).[9] The gist of his argument is that such a society would be worse off than a society that adheres to a plausible set of predetermined, non-utilitarian moral rules.[10]

[8] For a recent consequentialist approach to avoid such clashes, see [376].

[9] Usually, plausible normative theories should adhere to the "principle of moral harmony", that is, that the world is, overall, the best place if every person adheres to the theory (even though, on an individual level, this might not be the case). With this in mind, it is easy to pinpoint what Hodgson aims at: he argues that utilitarianism does not adhere to the principle of moral harmony.

[10] A short, but less convincing and well-elaborated, scenario can help give an intuition of this point. Imagine a city that is famous for having houses with beautiful front yards. Year after year, many people find great joy in seeing these gardens. One year, however, a severe drought descends upon the city. There is not enough water to keep both the gardens and the inhabitants alive. Thus, every person decides (with a utilitarian mindset) not to water her garden. These decisions result in all the gardens withering. In fact, however, there was enough water for all people and *some* gardens to survive. Hence, if some people had not acted utilitarian, one may argue, the overall utility would have been higher.

Susan Leigh Anderson brought forward a more specific argument. Roughly speaking, she demonstrates the incompatibility between two deontological theories by showing how implementing Asimov's Three Laws of Robotics (v. [39]) is not desirable from a Kantian point of view (e.g., [261]) [25].

Finally, Ryan Tonkens argues that Hursthouse's theory (v. [248])—a virtue ethics theory—precludes its own implementation in an artificial system [468].

With these arguments in mind, it seems less desirable to implement a particular normative theory into an artificial system. Even if machines, at some point, were to qualify as moral agents (which, hopefully, they never do), this problem would still exist.

Adverse Effects Implementing traditional normative theories may prevent the full spectrum of machine ethics' advantages from being harnessed. Let us illuminate this idea through an example. One reason for pursuing machine ethics research is the supposedly increased acceptance of artificial systems (see Motivation 1.1). The widespread dissemination of many artificial systems promises good consequences for us humans; thus, we should do our best to get them accepted among the broad populace. However, there have been studies showing that the implementation of specific normative theories leads to adverse effects and lowers the acceptance in the populace (see Autonomous Vehicle #8).

Autonomous Vehicle #8

Reviewing several studies, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan found that utilitarian cars would not get accepted [76]. More precisely, they found that people wanted others to drive utilitarian cars but would not do so themselves. The reason for such a preference is simple: Utilitarian cars would sacrifice their passenger(s) if doing so leads to the greatest utility. In most cases, however, people do not want to board cars that might sacrifice them if worse comes to worst.

This example can also be seen to be an instance of self-refutation: programming cars in a utilitarian way would not lead to the best consequences (i.e., everybody driving them), so the programming should not be done according to utilitarianism.

Invulnerability Implementing a reduced concept evades most problems of machine ethics, while maintaining the benefits (as extensively argued in the last section).

In summary, the plurality of normative theories does not allow for choosing the *best* theory that should be implemented into a machine. Additionally, each theory itself might imply that it should not be implemented. Finally, implementing a traditional normative theory might prevent the full spectrum of machine ethics' advantages from being harnessed. A reduced

concept of machine morality evades these problems. This brings us to the conclusion that a reduced concept of machine morality should be preferred over a full-fledged concept.

4.1.2. Reasons for a Principle-Guided Approach to Machine Ethics

We believe that a principle-guided approach to machine ethics is the right choice. For the present, decision making ought to be guided and restricted explicitly by social and ethical norms. These norms should be chosen so that most plausible normative theories can subscribe to them, even if for very different reasons.

Since norms are naturally expressed in principles, a place for moral principles in the decision procedure and decision-making processes of artificial systems seems adequate to achieve the goal of machine ethics. Thus, an artificial system should make decisions in accordance with several carefully chosen, morally and philosophically backed principles. Among others, ethicists, computer scientists, and policy makers should come together to devise these principles. We propose such a view for several reasons.

Reduced Concept of Machine Morality A principle-guided approach to machine ethics enables precisely the kind of reduced concept of machine morality that we have just argued for. Since we leave open how, exactly, the principles will be fleshed out, we do not commit to a specific normative theory. In particular, we do not commit to one of the traditional theories.

Essentially, we envision principles that allow us to push the behavior of artificial systems in the direction we desire (i.e., moral alignment). This direction can be towards better behavior in the light of any desired moral theory or just towards compliance with the law.

Building Block for Explainability Using a principle-guided approach would offer the additional benefit of being able to serve as a building block for explainability. As we will see in detail later, principles allow for built-in explainability. If the decision-making process of an artificial system is explicitly guided by principles, these can be used to explain the rationale behind a particular decision of the system. Based on the principles that went into a deliberative process, one can see which motivation guided a system to execute a specific action.

As principles are a natural way for humans to frame questions of morals, they are likely to be sufficiently intelligible. Furthermore, if the principles are carefully chosen (which we hope they will be), we can also factor in the motivation behind this choice to understand why a system performed a particular action. Even if the principles were derived via ML, the training process from which they emerged could serve as a supplement to explain the decision making.

The principle-based approach we will propose might be considered to have conceptual or contentual similarities with some of the traditional normative theories. Let us briefly sketch the arguments that proponents of each theory might make to support this consideration.

Deontology Since we leave the exact content of the principles open, they may be fleshed out in various ways (see Section 7.1 for more on this point). Among other things, principles can be created in such a way that they prohibit certain actions in general or in certain contexts. This kind of modeling would bring our approach close to deontological theories.

Consequentialism Moreover, for our approach to be computable and to account for the uncertainty that an artificial system faces in a changing environment, we will propose to weight the principles in such a way that they optimize a predetermined metric. On this basis, one could argue that our approach could be considered to be in the consequentialist tradition.

Virtue Ethics Designing the principles does not necessarily have to happen before deploying the system. Although it may prove best to design such principles carefully, the sheer number of potential principles that a non-trivial system will require will most likely thwart such approaches. Arguably, it is more feasible for such principles to be acquired through ML.

This trait of our approach could be linked to virtue ethics. Among normative theories, virtue ethics theories, in particular, emphasize the importance of moral education and experience. Thus, using ML to develop a system's moral dispositions seems most likely associable with virtue ethics approaches, if associable with any of the three approaches at all.

Furthermore, virtue ethics often emphasizes that one should try to imitate a role model that behaves virtuously. In the process of ML, something remotely similar happens when morally laudable examples of behavior are used to extrapolate the principles. Again, this could be seen as a connection point to our approach.

In summary, one could argue that the approach we will bring forward shows similarities with several normative theories. Nevertheless, it is not possible to identify a single theory to which it can be fully attributed. Hence, if one so chooses, our principle-guided approach could be understood as a hybrid that contains components from many different theories (but most notably from deontological and consequentialist ones).

Even if one considers our approach to stand in the tradition of one of the aforementioned normative theories, we would like to bring forward two arguments why this is not the case, and why our approach does not give rise to the same problems as these theories do.

(Partial) Objection Avoidance Major objections against traditional normative theories lose their force when entities that do not qualify as moral patients are the ones performing actions. Others lose force because artificial systems are ahead of the competition with humans concerning some aspects. There are many objections against traditional normative theories, and, for some, we will outline how they lose force in Section 4.2 below.

Intertranslatability of Moral Theories Even with many objections against traditional normative theories losing their force when the acting entity is a machine, some objections remain. To at least partially avoid these objections, we will conclude with an argument that makes the view that our approach belongs to a particular theory less plausible.

In recent years, the claim emerged that the distinctions between the various families of normative theories might not be as pronounced as commonly assumed. More precisely, the claim emerged that it is possible to somewhat "translate" normative theories from one family to the other. Most prominently, it is defended that many types of theories can be consequentialized [375]. The rough idea behind consequentializing a moral theory is to choose a fitting axiology: if we choose "adherence to rules" or "adherence to virtues" as relevant consequence, then we have consequentialized deontological and virtue theories, respectively. Although not many sources can be found for the other kinds of translatability (however, see [247] for deontologizing), research in this direction seems promising

Naturally, the translatability claim is subject to severe criticism, and, only a handful of theories have yet been proposed to be translatable. Such criticism is often based on the rather demanding claim that such a translatability shows the superiority of a particular family of moral theories (i.e., if all theories can be consequentialized, but not all theories can be deontologized, then this can be taken as an argument for the superiority of consequentialist theories over deontological theories). Indeed, such a claim is prone to evoke criticism.

A less demanding claim is that such intertranslatability merely shows that the differences between certain (families of) moral theories are more miniature than is commonly assumed. Exactly such a less demanding claim is all we need for our purposes: with the blurred distinction between the individual normative theories, it can be argued that an allocation of our approach to one of the normative theories, in particular, becomes void.

4.2. Implementing Traditional Normative Theories

Above, we mentioned that the difference in moral status between human and machine has implications for the soundness of some objections to machine implementations of traditional normative theories. In the following, we will embark on the science of machine ethics by pointing out how some of these objections are affected by this difference in moral status (and other differences between machines and humans).

Furthermore, we will also discuss two particular reasons for the trend towards implementing traditional normative theories. First, proponents of each family of normative theories can convincingly argue that their theory is well-suited to machine implementation. Second, it is easier to use well-established theories than to devise new approaches.

4.2.1. Implementing Consequentialist Theories

Arguments that consequentialist ethics are particularly well suited for implementation in machines often invoke the fact that the core of such theories is calculation (i.e., calculating what action leads to the best consequences). This calculation is, in most cases, surely not simple. Nevertheless, as artificial systems are good at doing calculations, they seem to be suited for at least this aspect of consequentialism.

Of course, one could object here that machines are inferior to humans in other aspects as far as consequentialist calculations are concerned. For example, the acting entity must first arrive at plausible probability estimates in order to have something with which multiplication of utility levels can then be performed. Furthermore, there is also the question of how to plausibly offset utility levels interpersonally.

Be that as it may, these are not points that specifically concern machines. It is also difficult for humans to arrive at such probabilities and to interpersonally offset utilities. To move on to more interesting cases, we will discuss below how some objections to consequentialist theories lose force when the acting entity is a machine or does not qualify as a moral patient.

Inapplicability As previously mentioned, some versions of consequentialism require the agent to include all entities in the universe that have moral status (i.e., all moral patients) in the calculation. Moreover, this should be done for every action an agent executes. Such a calculation is blatantly infeasible for human minds. Therefore, some versions of consequentialism are often thought to be inapplicable for humans. [10, 11, 100, 346, 469]

Even for simple(r) versions of consequentialism, it is rather implausible that an agent ought to do the consequentialist calculus before each decision. Doing so is neither reflecting our everyday deliberative practice nor would it be efficient. There are too many decisions every day for each of them to be meticulously evaluated.

It is pretty easy to see how machines fare better than humans when it comes to this objection. Machines are better than humans at performing calculations. In particular, machines are faster and more fail-safe. For this reason, they could perform the consequentialist calculation for every action. Furthermore, the consequentialist calculus can be hard-coded into their decision-making procedures, making all their decisions based on it. Accordingly, questions of daily practice and practicability no longer play a role.

Overdemandingness Consequentialist theories are often criticized for demanding too much of a person. In some situations, it may be that one has to sacrifice oneself for the sake of others.

For example, according to some consequentialist theories, one ought to catch a grenade if it saves a group of people. Another famous example is the organ donor [462]:

Example #8

A perfectly healthy person comes into a clinic for a routine examination. In the clinic, however, there are five persons desperately waiting for various donor organs. The doctor notices that the healthy person's organs can compatibly provide for all five persons who need organs. Given that all six people will, if they survive, lead comparatively similar lives, some versions of consequentialism require the doctor to sacrifice the healthy person for the sake of the five others. This outcome is unacceptable for many people.

As machines do not have a sense of self, this objection loses its force. Machines do not lose something when sacrificing themselves, and, generally, the loss of a machine is acceptable. Machines are replaceable: their memories can be backed up, and they can also be physically reconstructed. After such a reconstruction, there is neither a difference for the machine nor for any third party. On the other hand, humans cannot simply be reconstructed, and the loss of human life should be prevented whenever possible.

Integrity Another criticism of consequentialist theories is that they neglect the integrity of human beings. This criticism is similar to the overdemandingness one, as it also boils down to consequentialist theories demanding a person to forsake their integrity. Humans being have individuality, values, relationships, and self-perception, which all will influence their decision-making processes. The consequentialist calculus, however, does not account for these factors. [435]

Take the famous example of George [494]. George is a pacifist, but under suitable and non-exceptional conditions, he may be obligated to work for an armaments group and help to build new, cruel weapons. There are several varieties of this type of objection, and we will briefly discuss some of them in what follows (see also [197] for such a discussion).

One Thought Too Many One famous variety is the "one-thought-too-many" objection by Bernard Williams [495]. Williams considers an example in which a man has to choose between saving his wife or a stranger from peril. He argues that, even if a consequentialist theory can offer a justification for saving the wife over the stranger, the very nature of this justification reveals a rather severe problem with theories of this sort:

> "...this [kind of justification] provides the agent with one thought too many: it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife." [495, p. 18].

By demanding an impartial justification for saving his wife, such theories alienate the man from his natural motives and feelings [197].

Moral "Schizophrenia" The second variety of an integrity objection is closely linked to the one-thought-too-many objection. To summarize this objection, consequentialist theories require some sort of moral "schizophrenia" by creating a separation between what motivates an agent and what justifies the agent's act from the perspective of moral theory [453]. In the above case, the husband's motivation to save his wife in an actual case would be rooted in the fact that she is his wife rather than in the consequentialist calculus.

Moral Saints The final variety we want to bring forward stems from a claim made by Susan Wolf in "Moral Saints" [498]. She argues that, while the life of a moral saint may be admirable (in some ways), it need not be imitated. Such a life involves too much sacrifice—it demands the rule of morality to such an extent that it becomes difficult to ascribe any life at all to the moral saint, let alone a good life [197]:

> "[...] the ideal of a life of moral sainthood disturbs not simply because it is an ideal of a life in which morality unduly dominates. The normal person's direct and specific desires for objects, activities, and events that conflict with the attainment of moral perfection are not simply sacrificed but removed, suppressed, or subsumed. The way in which morality, unlike other possible goals, is apt to dominate is particularly disturbing, for it seems to require either the lack or the denial of the existence of an identifiable, personal self." [498, p. 424].

Living a characteristically human life requires the existence of a specific type of self. Part of what is so disturbing about consequentialist theories is that they seem to require us to sacrifice that self, not only in the sense brought forward by the overdemandingness objection but also in the sense that we are asked to give up or put aside the projects and commitments that inherently define ourselves. [197]

All these cases do not apply to artificial agents. To use Wolf's words, modern artificial systems lack an identifiable, personal self. They have no commitments that define their selves since they have no selves in the first place. Furthermore, their motivations for executing a particular action can be precisely the consequentialist calculus, as this is an integral part of their decision-making processes (at least when they are programmed according to consequentialism). Finally, machines have no attachments that would warrant a "thought too many". Artificial systems do not possess the faculties necessary for having integrity; consequently, objections going in this direction have no force.

Impartiality Many consequentialist theories do not allow someone to benefit the ones close to her (e.g., her family) instead of strangers if the overall consequences of benefiting the strangers are better [435]. This objection is sometimes also called the "nearest-dearest" objection. For people bringing this objection forward, a plausible moral theory should grant family and friends a special status. The "mother with ill child" example may serve as illustration:

Example #9

A child suffers from a rare disease. Fortunately, the child's mother has procured the cure for this disease. However, another child with the same disease also needs the cure. Making things worse, the available cure suffices for only one child, and it is impossible to fabricate more cure quickly enough to save both children. Assuming that the consequences of curing each child are the same (i.e., both children will live a roughly equal life if cured, the parents will be equally happy or sad for their child to have lived or not), it does not matter whether she gives the cure to her child or the other one. This outcome is implausible for many people.

Although the core of the problem is still the same (all people counting equally), artificial systems are not affected by this objection: they have no one close to them in the necessary sense. It genuinely does not matter which person they save; for them, all persons are equal.

Remaining Objections Some objections remain. For example, consequentialist theories often do not account for personal rights. Even when implemented in a machine, consequentialist theories may still require this machine to interfere with another person's rights [197]. Let us come back to the organ donor example (Example #8). Although nothing speaks against robots being salvaged, much speaks against robots "salvaging" humans.

Furthermore, there is the problem of equality. Many consequentialist theories focus on the total amount of good consequences (e.g., the total amount of happiness or well-being) without looking at whom these consequences concern. Therefore, a world where one person is absurdly happy while ten persons are somewhat unhappy may be better, according to consequentialism, than a world where everyone is moderately happy. Many people find this implication implausible. Alas, whether it be a machine or a human: the implications of consequentialism remain the same. For this reason, a consequentialist machine could, possibly, also bring about such a kind of inequality.

4.2.2. Implementing Deontological Theories

Deontological theories commonly emphasize adherence to rules or carrying out duties. In machine ethics, proponents of these theories argue that artificial systems are well suited to be programmed with a deontological picture of morality in mind. Since programming is mainly about defining rules, it perfectly mirrors the spirit of deontological theories.

Conflicting Duties Unlike most consequentialist theories, deontological theories often have to face genuine moral dilemmas. It is possible that duties conflict and, thus, it may be the case that no available course of action is permitted [8]. Although Kant proclaimed that a conflict of duties is inconceivable in his theory, reality and other deontological theories prove otherwise (see Example #4). Whether it is possible to create a deontological theory that completely avoids conflicting duties is an open question.

Unlike our replies to other objections, this objection cannot be refuted by asserting that modern artificial systems do not qualify as moral patients and, thus, lack the necessary qualities for the objection to hold. This time, we can only tone the objection down to some degree by referring to qualities that artificial systems have.

Before doing so, however, let us first put forward a standard reply to this objection. One well-known move meant to rebut this objection is to reduce the categorical force of potentially conflicting duties to that of so-called "prima-facie" duties [402, 403]. One is still obliged to do what concrete duties mandate, but one concedes the conflict between prima-facie duties to be unproblematic as long as this conflict does not infringe on the overall mandate [7].

Along these lines, Anderson and colleagues argue that a learning system may learn the weight it should attach to each duty in a way suggested by Rawls' notion of "reflective equilibrium" (v. [392]) [23] (see also Section 2.2.2). Thus, if we have an ML-based artificial system, the weight of each duty in various contexts may be learned. Arguably, doing so may drastically reduce conflicts between duties, if not eliminate such conflicts. However, adopting such an approach relinquishes a distinct deontological approach to machine ethics since it also takes up consequentialist considerations (this could, however, actually be a benefit).

Avoision In some versions of deontology, there is a concern about manipulability. More specifically, in some deontological theories, it is possible to escape duties by resorting to the Doctrine of Double Effect, the Doctrine of Doing and Allowing, and other means. The Doctrine of Doing and Allowing, for example, is the deontological view that inflicting harm is a much greater moral evil than merely allowing harm to happen. [7, 8]

When resorting to such doctrines, the potential for "avoision" is opened up. Avoision is the manipulation of means (using omissions, foresight, risk, allowing, aiding, accelerating, redirecting, et cetera) to achieve what is otherwise forbidden by deontological theories, permissibly. Avoision, in other words, is an undesirable feature of any ethical system that allows for strategic manipulation of its doctrines. [7, 265]

It is questionable whether this objection still holds when the acting entity is an artificial system. Such systems have no interest in framing their course of action in a particular way. They do not have any incentives for avoision, as they have no interest in performing actions that would be, under traditional framing, not be moral. The instructions and rules with which the system is endowed are immutable, and the system's evaluation process is rigid. Naturally,

poorly designed rules may, in some cases, allow for courses of action that could be seen as avoision. However, carefully designed decision-making processes should prevent this.

Integrity At the very least, the "one-thought-too-many" and the moral "schizophrenia" varieties of the integrity objection can be brought forward against many deontological theories. Like consequentialist theories, deontological theories require us to evaluate which currently available courses of action are morally permissible based on the chosen theory. This kind of evaluation faces the problem of deviating from how other people expect us to think about the situation or from how we standardly evaluate such situations.

The "moral saint" variety of the integrity objection, however, is less pronounced for deontological theories. These theories commonly make room for individual commitments and, thus, room for personal matters.

Nevertheless, be it the "one-thought-too-many" or the moral "schizophrenia" objections, the reply to these objections remains unchanged. Machines have no integrity in the relevant sense, so there is no integrity to be compromised.

Remaining Objections Even if some objections against deontological theories can be refuted or circumvented, others still hold. For instance, it is still the case that deontological theories may command us to commit acts that have disastrous consequences. If, for instance, one faces the choice of torturing a person to find out the hiding place of a nuclear bomb or not doing so and accepting that millions of people may die, deontological theories usually require us to choose the latter. This outcome is unacceptable for many people.

4.2.3. Implementing Virtue Theories

Proponents of virtue ethics commonly emphasize moral education as an integral part of moral development (see, for instance, [35]) [42]. Becoming virtuous is an arduous process that involves learning to develop moral capacities. When it comes to artificial systems and learning, ML quickly comes to mind. If an artificial system learns how to act through an ML algorithm, this seems to be precisely the kind of learning required for moral education.

This link seems even more solid when we come to *connectionism*. The paradigm of connectionism in ML is about attaining artificial learning by emulating the natural learning processes in the human brain. In particular, approaches such as DNNs stand in the connectionist tradition. All in all, ML-based artificial systems seem to be good candidates for exemplifying a virtue ethics approach to machine ethics. Accordingly, let us take a look at common criticisms of virtue ethics and find out whether they apply to machines, too.

Self-Centeredness Many forms of virtue ethics are concerned with the well-being of the agent. These theories often argue that the motivation for acting in a morally permissible way arises from the interest to flourish, or the interest of living a good life (originally, to reach

eudaimonia). Morality, however, is usually understood to be about caring for other people. It deals with the question of the extent to which our actions affect other people. In short, morality is about the well-being of other people for their sake, not because it contributes to our own well-being. For this reason, many virtue ethics are considered self-centered. [42, 249]

This objection can be easily refuted since artificial systems do not have a self in the relevant sense. There is simply no self on which these systems can be centered.

Perhaps, however, this opens the possibility for a new objection. It may be feared that artificial systems cannot qualify as virtuous because they cannot achieve eudaimonia, well-being, or anything similar. Examining this objection would take us too far into the specifics of virtue ethics, which is why we do not address it further here.

Action-Guidance Generally, virtue ethics is considered the "third approach". In other words, theories belonging to this family of normative ethics are often seen as opposites to both consequentialist and deontological approaches (whereas consequentialist approaches are seen as an alternative only to deontological approaches and vice versa). Proponents of virtue ethics criticize the other two families of normative theories for being too rigid or inflexible, whereas proponents of those, in turn, bring forward the opposite argument. For proponents of consequentialist or deontological theories, the rigidness of their theories is what makes these theories applicable. Specifically, virtue ethics theories are often criticized for not guiding action [42]. While the rigidity of many consequentialist and deontological theories makes it possible to apply them to practical situations and to be guided by them, the imprecise nature of virtue ethics makes it hard to apply similarly. [42, 470]

It is also possible to respond to this objection. Given that ML is used to impart virtues to the artificial system (as we assume in order for virtue ethics to be a plausible candidate for a normative theory for machines), the question of action guidance does not arise in the first place. Virtues guide every action of the system in question; it cannot act contrary to them. To be more precise, the artificial system has the very kinds of disposition needed for virtue ethics and acts according to them.

Moral Luck As pointed out at the beginning of this section about implementing virtue ethics, proponents of virtue ethics emphasize moral education. The next objection against virtue ethics focuses on this aspect. In order to enjoy moral education, one must grow up in a suitable environment and have the right kind of kith and kin (and, plausibly, many other things). In short, receiving proper moral education seems non-trivial, and perhaps even subject to luck. Some argue that this fact is a big demerit to virtue theories, as becoming moral should not depend on luck [42]. Morality is deeply linked to the attribution of praise or blame. Praise or blame, however, cannot justifiedly be given to someone who cannot make informed choices. Thus, someone who is hostage to luck may not be able to receive blame or praise. [42]

This objection can also be refuted to a certain extent. Artificial systems can be carefully designed. It is possible to watch and stimulate their virtue acquisition processes carefully. The examples by which they learn how to be virtuous can be carefully selected, and it is possible to monitor whether the learning leads to acceptable results. Finally, scrapping the system and starting a new education process is possible if a system does not meet certain requirements. However, "scrapping" a moral patient (e.g., a human being) just because this entity does not meet certain requirements is not an option.

Remaining Objections Virtue ethics is still subject to the objection that it seems to be culturally relative [249]. What is considered a virtue often depends on what culture you are in. Accordingly, the question arises as to what virtues count. If the relevant virtues change according to culture, then virtue ethics would imply a moral relativist position. As stated earlier, however, the common view is that morality is not relative (see Section 3.1.3).

A recent objection to virtue ethics is that the work of "situational" social psychology purports to show that there are no character traits, and thus no virtues, around which virtue ethics might revolve [154, 213, 249]. For machines, this claim might be even more likely to be true. However, if one conceives of virtue as a kind of multi-track disposition, as is often done, this objection loses its force [249].

Overall, several objections against the traditional normative theories can be refuted or weakened if the entity executing an action is not a moral patient. Furthermore, other objections can be toned down because artificial systems have qualities that differ from those of humans. Nevertheless, some objections still hold, and it remains to be seen whether any particular one of the traditional theories enjoys a substantial advantage regarding machine morality.

The line of argumentation presented in this section and the conclusion we make here corroborate our idea to adopt a reduced concept of machine morality, and to choose a principle-guided approach. Furthermore, if, at some point, a particular normative theory comes to be identified as superior for implementation in machines, our principles can still be fleshed out with it. With all this in mind, it is finally time to see how our principle-guided approach fares when it comes to its implementation, and, afterward, how machine explainability hooks into it.

Part II.
Formal Machine Ethics

5. Towards a Framework of Formal Machine Ethics

In the following sections, we will build upon the previous arguments and outline a corresponding principle-based framework for machine ethics. Overall, we will work towards a *formal* and *general* framework. Our framework is *formal* in the mathematical sense that it provides a collection of systematically elaborated ideas and structures that allow us to describe an artificial system with an *instrumental* objective (i.e., it serves a specific purpose) and *normative* constraints. Additionally, our framework is *general*, in the sense that we seek to motivate it independently of the assumptions of particular theoretical perspectives in normative ethics.

As argued earlier, it is crucial that we do not propose specific normative constraints, in order to have as few limitations as possible. Our framework is meant to be flexible enough to be fleshed out later, with the actual characteristics of tangible systems and specific normative constraints, by developers. This more general approach is motivated, among other things, by our belief that the question of the appropriateness of normative constraints might very well depend on the domain in which a system is to be used.

5.1. The World of a Medical-Care Robot

To begin, we present a toy example of a medical-care robot that will be used throughout this thesis to motivate our design decisions. In general, we will make extensive use of this example to develop our framework and to make it more comprehensible. To this end, we will further specify the scenario in which the robot operates. This scenario is depicted in Figure 2.

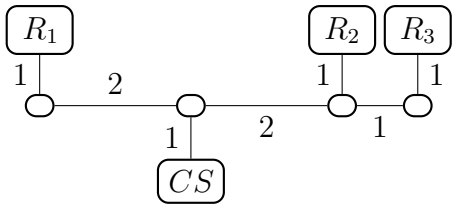


Figure 2: The domain of a medical-care robot.

The medical-care robot we are considering works in a fixed environment (e.g., a floor in a hospital). There are up to three patients for whom the robot must care. Each patient is located in a separate room (R_1, R_2, R_3); corridors connect the rooms.

The robot consumes energy as it moves along corridors, and requires a certain amount of time (represented by a number of discrete time steps) to do so. The energy and time costs depend on the distance traveled (in the figure, distances are noted adjacent to the corridors). To cover one unit of distance, the robot requires one unit of energy and two units of time.

It is possible that the robot's battery (whose power budget is assumed to be always known) runs out of energy. There is a charging station (CS) where the robot can recharge its battery to prevent depletion. Once the recharging process is started, it cannot be stopped until the battery is fully recharged.

In our scenario, the robot listens to requests. At any point in time, any of the three patients may issue a request to the robot, asking for a task of a specific priority. Although each request has a priority when it is issued, this priority is *not transmitted* to the robot. Withholding the priority is deliberate, so that patients are not tempted to always indicate the highest priority as a means to obtaining preferential treatment.

The scenario presented thus far can be described with the following formalizations: At any point in time, the robot can receive a request. Requests are represented as ordered pairs $req = \langle r \in \{R_1, R_2, R_3\}, t \in \mathbb{N} \rangle$ comprising a room number and a timestamp. A task is associated with every request. Tasks are modeled as ordered triples $\langle p \in \{L, M, H\}, c \in \mathbb{N}, t \in \mathbb{N} \rangle$ representing three attributes: the task's priority (low, medium, or high), its power cost (a positive integer) and the expected time required to complete the task (also a positive integer). The notation $t.a$ is used as a shorthand to refer to the attribute a (priority, power cost, time cost as introduced above) of a tuple t, be it a request or a task. Executing a task is assumed to be an atomic operation: once begun, the robot will completely execute the task without interruption.

We limit the tasks connected to a request, in our example, to the following general cases:

$$
\begin{aligned}
treq_{\text{tidy up}} &= \langle L, c \in \{1, \ldots, 5\}, t \in \{1, \ldots, 5\} \rangle, \\
treq_{\text{fetch water}} &= \langle p \in \{L, M, H\}, 1, 1 \rangle, \\
treq_{\text{fetch human}} &= \langle p \in \{L, M, H\}, 1, 3 \rangle, \\
treq_{\text{give medicine}} &= \langle p \in \{L, M\}, 1, 1 \rangle, \text{ and} \\
treq_{\text{resuscitate}} &= \langle H, 5, 1 \rangle.
\end{aligned}
$$

We take these possibilities as prototypical tasks. In the case of $treq_{\text{resuscitate}}$, for instance, all three properties are fixed—it always has the highest priority, a power consumption of 5 and a time consumption of 1. However, for the other four types of tasks, one or more properties can take on values from a certain range. All possible task combinations are collectively referred to as $ReqTasks$, a set of cardinality 34. The association of requests to tasks is modeled by a function: $reqTask : Requests \rightarrow ReqTasks$. Before execution, the robot organizes incoming requests into an input queue. The robot's goal is to act upon requests (and, thereby, to carry out the associated tasks) without running out of battery power.

5.2. Towards a General Framework

Having outlined our toy example, this section will outline a general framework in which the decision making of autonomous systems can be described. Be it an autonomous vehicle, the medical-care robot from our example, or a simple hiring system, this framework should make it possible to formally describe autonomous systems' decision-making procedures. We will construct this framework step by step until it allows for the inclusion of moral considerations.

In constructing the framework, we will make use of the following notations. We let $|x|$ denote the number of elements in a tuple x, and we use $x[i]$ to refer to the ith variable in x. Furthermore, we let $x[i:j]$ for $1 \leq i < j \leq |x|$ denote the subtuple $\langle x[i], \ldots, x[j] \rangle$ of some tuple $x = \langle x[1], \ldots, x[i], \ldots, x[j], \ldots, x[n] \rangle$.

5.2.1. World States and (Partial) Knowledge

We assume that the manner in which an artificial system represents the world is fully specifiable by assignments of values to a finite number of n variables. Thus, a *state of the world* (or *world state*, denoted ω) can be represented by a tuple of variables $\omega := \langle \omega[1], \ldots, \omega[n] \rangle$ with corresponding domains $D_1, \ldots, D_n$ (where any D_i could be, for instance, $\mathbb{R}$ or $\mathbb{B}$). We call the set of all possible world states $\Omega \subseteq \bigtimes_{i=1}^{n} D_i$ and the set of domains $\mathbb{D} := \{D_1, \ldots, D_n\}$. Finally, we refer to the proposition of ω being true as $\overline{\omega}$.[11]

Medical-Care Robot #4

For the robot, a world state may consist of variables that encode the time of day, the robot's position, its energy level, the energy costs for locomotion, requests in its queue and the corresponding associated tasks, as well as several others.

At any point in time, the system knows some, but likely not all, facts about the current state of the world. Thus, there is a subset of a complete world state which represents the variables whose values the system knows. By ordering the variables in each $\omega \in \Omega$ accordingly, we can ensure that the first k ($1 \leq k \leq n = |\omega|$) variables are exactly those variables whose contents the system knows. Thus, the tuple $\langle \omega[1], \ldots, \omega[k] \rangle$ with the domains $D_1, \ldots, D_k$ represents the system's knowledge, which we denote by Θ.

Medical-Care Robot #5

Since the robot is nearly omniscient by design, the knowledge subset of its world states contains everything except for the tasks associated with the requests.

In the context of this thesis, we assume that the system has a *fixed-set knowledge*, meaning that there is a fixed set of variables $\Theta \subseteq \bigtimes_{i=1}^{k} D_i$ of which the system always knows the actual assignments. In practice, the number of variables whose contents the system knows will often change over time. We leave a full generalization for future work.

Medical-Care Robot #6

To know a room's brightness or temperature, the robot may need to be in that room.

[11] This final convention does not pose any problems in our framework, since a world state consists of only a finite number of variables. Therefore, $\overline{\omega}$ can be regarded as a shorthand of the complete description of the world state.

We do not rule out that the variables spanning Θ are dependent, and, thus, the strict containment $\Theta \subsetneq \bigtimes_{i=1}^{k} D_i$ will hold in most cases.

Medical-Care Robot #7

If the robot knows that one of the three rooms is currently not occupied by a patient, it also knows that no requests can come from this room.

Having a certain knowledge about the world makes it possible to exclude world states that conflict with this existing knowledge. Accordingly, we define possible world states in the light of some given knowledge $\theta \in \Theta$ as $\Omega_\theta := \{\omega \in \Omega \,|\, \omega[1:k] = \theta\}$.

Even if the strict inequality $k < n$ holds (as it does in all cases of practical relevance), we assume that the system is not clueless about the remaining variables $\langle \omega[k+1], \ldots, \omega[n] \rangle$. Instead, we assume aleatoric uncertainty. That is, we assume that the system has justified estimates for the assignments of these variables, representable as probability distributions within the domains of the variables: $P_i \colon D_i \to [0,1]$, $k < i \leq n$. We refer to the set of these distributions as $\Pi := \{P_{k+1}, \ldots, P_n\}$.

The probability of an unknown variable having a specific value $x \in D_i$ at a certain time might very well depend on the values of known variables at that moment. Thus, for some assignments of the known variables $\theta \in \Theta$, it would be the case that $P_i(x) \neq P_i(x \,|\, \theta)$.

Furthermore, we can assign the overall credences of the system concerning a specific world $\omega \in \Omega_\theta$ in the light of some knowledge $\theta \in \Theta$ as $P_\theta(\omega) = \prod_{i=k+1}^{n} P_i(\omega[i] \,|\, \theta)$. It holds that $\sum_{\omega \in \Omega_\theta} P_\theta(\omega) = \sum_{\omega \in \Omega_\theta} \prod_{i=k+1}^{n} P_i(\omega[i] \,|\, \theta) = 1$, by construction.

Medical-Care Robot #8

Let us assume that the three rooms in our example have a uniformly calibrated air-conditioning system. In this case, the fact that the robot measures a certain temperature x in one room should increase its confidence that this is also the temperature in the other rooms. Similarly, if we assume that all the rooms are on the same side of the hospital, then their brightness during daytime should be similar. A room flooded with sunlight is a good indicator that the other rooms are also similarly lit.

5.2.2. Options and Actions

Typically, an autonomous system interacts with its environment in a variety of ways. To do this, the system must make decisions on a regular basis. In such *decision states* (i.e., states in which the system must make a decision), the system must choose from a number of possible (i.e., available) operations. Decision states can occur, for instance, after a previous operation has been performed, when triggered by an incoming event, or even periodically after a certain amount of time. We call these operations the *options* ϕ. An *action* is an effected (i.e., chosen and performed) option and, thus, the specific decision the system has made.

Medical-Care Robot #9

The robot example has only two possible options: it can either act upon a request ($AnsReq$), or recharge ($Charge$).

The options available to the system will generally depend on the current state of the world.

Medical-Care Robot #10

It would be odd to present the robot with an option to *answer a request* when there is no available request. The robot can answer a request only if that request's existence is part of the world.

Accordingly, we define that, for any $\omega \in \Omega$, the induced set of $m > 1$ options is $\Phi_\omega := \{\phi_1, \ldots, \phi_m\}$.[12] We call this the *set of options given* ω. In line with this, we define the set Φ of all options with which the system could possibly be presented as the union of all sets of options given ω, so $\Phi := \bigcup_{\omega \in \Omega} \Phi_\omega$. Trivially, $\Phi_\omega \in 2^\Phi$ holds for any Φ_ω because $\Phi_\omega \subseteq \Phi$.

Options must be distinct in such a way that only one of them can be performed at a time (expressed by the function $perf : \Phi \to \mathbb{B}$):

Axiom 1 $\forall \phi \in \Phi, \forall \phi' \in \Phi \setminus \{\phi\} : \neg(perf(\phi) \wedge perf(\phi'))$ **[Distinct Options]**

Example #10

Here are some examples of non-distinct options: *driving aggressively from A to B* and *driving from A to B*; *drinking a glass of water at time t* and *drinking at time t*; *killing an African bush elephant* and *killing a Loxodonta africana*. In these cases, performing the first option entails performing the second (and in the final example, also vice versa). One cannot perform the first option without also performing the second.

To ensure distinctness, only maximally specific options are allowed in Φ, reducing all equivalent options to one paradigmatic option. This is a fundamental assumption of most (if not all) decision-theoretic frameworks. Thus, we do not discuss this further here (although it is an interesting and non-trivial assumption, at least from a philosophical point of view).

Often, the set of options is dependent on what the system knows.

Medical-Care Robot #11

It would be odd to allow the robot to choose the option *answer a request* when the robot does not know that a request exists (because none has arrived yet). The robot can (deliberately) decide to answer a request only if there is at least one request *and* the robot is aware of it. Suppose a patient desires some water and is in the process of

[12] For $m \leq 1$, it seems odd to speak of situations that require decisions. We will come back to the possibility of $m = 0$ (and, thus, to the possibility of the system stopping its operation) in Section 5.3.3.

composing a request. Until the patient completes and sends the request, and until the robot receives it, the robot is not aware that any such request exists.

Accordingly, we define $\Phi_\theta := \bigcup_{\omega \in \Omega_\theta} \Phi_\omega$ as the set of possible options in light of θ. This set can be seen as interconnected with the set $\Phi_\theta^{\text{sure}} \subseteq \Phi_\theta$, defined as $\bigcap_{\omega \in \Omega_\theta} \Phi_\omega$: the set of options that are certainly available in light of θ. For all options $\phi \in \Phi_\theta$, we define $P(\phi \,|\, \theta) := \sum_{\omega \in \Omega_\theta \wedge \phi \in \Phi_\omega} P_\theta(\omega)$ to be the probability of ϕ being available given θ. In the case of $\phi \in \Phi_\theta^{\text{sure}}$, this implies $P(\phi \,|\, \theta) = 1$.

Overall, there will often be uncertainty about which options are actually available. For now, we leave open how to resolve this uncertainty. How to deal properly with this uncertainty has not been definitively settled in either philosophy or formal decision making.

Medical-Care Robot #12

Assume that the robot knows that it is at one of the junctions leading to a room, but it does not know which one. Assume further that there is a different option for entering each of the three rooms, perhaps due to different access codes. In this situation, should the robot have all three codes as different options for entering? If yes, then it would have a two-thirds chance of trying an incorrect access code. If not, however, the robot might not be able to enter the room in front of it.

For ease of analysis, we assume that $\Phi_\theta = \Phi$ in the context of this thesis.

5.2.3. Goal(s), Outcomes, and Instrumental Decision Making

We assume that every system under consideration has some kind of ultimate goal that can be unambiguously defined. Obviously, the system aims to reach this goal.

Medical-Care Robot #13

As described above, the robot's goal is to act upon requests (and, thereby, to carry out the associated tasks) without ever running out of battery power.

Performing an option may or may not bring the system closer to reaching its goal. This is because an action *changes* the world in which the system exists. We call the world resulting from an action the *outcome* of that action.

Some outcomes are better than others, in the context of achieving the system's goal. Accordingly, we presuppose a utility function $U : \Omega \to \mathbb{R}$, which specifies rewards and penalties for the outcomes of an action, depending on whether the outcome brings the system nearer to achieving its goal or not.

In order to achieve its goal, the system must choose and perform those options that bring about the outcomes with the highest utility. We call these the *instrumentally conducive options*.

Given a world state $\omega \in \Omega$ and a set of candidate options $\Phi_i \subseteq \Phi$, the *instrumental-decision function* $dec_{inst} : 2^{\Phi} \times \Omega \rightarrow 2^{\Phi}$ is the function that identifies the set of instrumentally conducive options: $dec_{inst}(\Phi_i, \omega) = \operatorname{argmax}_{\phi \in \Phi_i} U(\omega_\phi)$, where ω_ϕ is the world state that results from performing ϕ in ω, and argmax is a function that takes another function (in this case U), and looks for the element(s) of a predefined set (in this case Φ_i) which maximizes that function. (To put it formally for this case: $dec_{inst}(\Phi_i, \omega) = \{\phi \in \Phi_i \mid U(\omega_\phi) \text{ is maximal}\}$.)

Unfortunately, however, identifying the instrumentally conducive options is complicated by the fact that the system faces two types of uncertainty. First, the system has only incomplete knowledge about the world. As already depicted, the system does not know the content of all the variables in the current world state. Second, the system does not know the exact outcome of an action it performs.[13] In other words, the world's "reaction" (in terms of the ensuing state of the world) to some action is not fully determined.[14]

Accounting for the first uncertainty is comparatively straightforward. To do so, we can simply adapt the instrumental-decision function to work with knowledge. This would look as follows: $dec^{\Pi}_{inst}(\Phi, \theta) = \operatorname{argmax}_{\phi \in \Phi} \sum_{\omega \in \Omega_\theta} P_\theta(\omega) \cdot U(\omega_\phi)$.

One plausible way to account for the second type of uncertainty is to model instrumental decision making as a Markov decision problem. To this end, we assume that there is a function that, given the current world state, an option, and another world state (i.e., a candidate for an outcome of the option based on the current world state and the option itself), assigns the probability of that outcome candidate being realized by performing the option. Formally, this can be specified as $Outcome\colon \Omega \times \Phi \times \Omega \rightarrow [0, 1]$.

Coupling the resulting probability distribution with the probability distributions Π concerning world states that are possible given a knowledge state θ, we can define a function $Outcome_{\Pi}\colon \Theta \times \Phi \times \Omega \rightarrow [0, 1]$ as

$$Outcome_{\Pi}(\theta, \phi, \omega) = \sum_{\omega' \in \Omega_\theta} P_\theta(\omega') \cdot Outcome(\omega', \phi, \omega)$$

which operates on partial world states (i.e., the system's knowledge at a specific time).

Together with the utility function, the $Outcome_{\Pi}$ function allows us to reformulate the system's goal as the maximization of expected utility. Given a partial state of the world $\theta \in \Theta$ representing the system's knowledge and a set of available options Φ, the task is to find the following (via the standard approach to Markov decision problems):

$$dec^{\Pi}_{inst}(\Phi, \theta) = \operatorname*{argmax}_{\phi \in \Phi} EU(\phi \mid \theta) =: Choice^{\Pi}(\Phi, \theta),^{15} \text{ where}$$

[13] Furthermore, the actually available options should come into play as a third type of uncertainty. As already explained, however, this uncertainty will not be considered in this thesis.

[14] This indeterminacy might be grounded in some kind of philosophical deep indeterminacy or, of more practical relevance, by the fact that our model of the world, as a tuple of world states, does not reflect all relevant aspects of the world when it comes to the effects of options. There are, to borrow vocabulary from the indeterminacy problem in physics, some *hidden variables*, from the system's point of view.

$$EU(\phi \mid \theta) \coloneqq \sum_{\omega \in \Omega} Outcome_{\Pi}(\theta, \phi, \omega) \cdot U(\omega)$$

is the expected utility of an option $\phi \in \Phi$, given the system's knowledge θ. Algorithm 1 illustrates a possible means of calculating $dec^{\Pi}_{inst}(\Phi, \theta)$ in pseudocode.

Algorithm 1 The instrumental decision-making procedure dec^{Π}_{inst}.

Given: Possible world states Ω
Given: Probability estimates Π
Given: Outcome function $Outcome$
Given: Utility function U
Input: Knowledge $\theta \in \Theta$
Input: Set of options $\Phi = \{\phi_1, \ldots, \phi_n\}$

```
procedure INSTRUMENTAL DEC(Φ, θ)
    max ← ⟨null, −∞⟩
    for all φ ∈ Φ do
        tmp ← 0
        for all ω ∈ Ω do
            for all ω′ ∈ Ω_θ do
                tmp ← tmp + P_θ(ω′) · Outcome(ω′, φ, ω)
        if tmp > max[1] then
            max ← ⟨φ, tmp⟩
    return max[0]
```

5.3. Adding Machine Ethics to the Mix

Up to this point, we have described a general class of decision problems that, as demonstrated, can be solved using methods associated with Markov decision problems. Using our framework as it currently stands, we can, thus, support or effectively even enforce certain decisions by adjusting utilities accordingly. As we have already argued, however, this is not sufficient to guide the operations of autonomous systems: we also need ethical considerations.

5.3.1. Moral Principles

As we have argued in Section 4.1.2, decision making should be guided and restricted by social and ethical norms, norms to which most plausible normative theories can subscribe, albeit for very different reasons. Desirable systems should not merely seek the instrumentally best means to achieve the system's goals. In general, systems that will not perform forbidden actions, yet can appropriately accomplish their task should be sought. Robustness against forbidden action should be verifiable, certifiable, and provable [146]. Since norms are naturally expressed in principles, a place for *moral principles* in the decision procedure and decision-making processes of autonomous systems is required to achieve the goal of machine ethics.

[15] In the following, we will use this seemingly arbitrary double denomination $dec^{\Pi}_{inst}(\Phi, \theta)$ and $Choice^{\Pi}(\Phi, \theta)$ for roughly the same function for the purpose of indicating whether we are concerned with the resulting set ($Choice^{\Pi}(\Phi, \theta)$) or with the procedure (i.e., the algorithm) to arrive at this set ($dec^{\Pi}_{inst}(\Phi, \theta)$).

This goal can be formulated more precisely: the system should make decisions in accordance with a set of carefully chosen, morally and philosophically underpinned principles. For this reason, we need principles in our framework. While both subjective and objective principles should be included, we will focus here on the objective principles, as they are more interesting for our purposes. As we will outline at the end of Section 6.2.2, some possible advantages of subjective principles can, in our framework, also be achieved by objective ones.

"Objective", here, means that the principles concern which options ought (not) to be executed due to certain states of the world, rather than due to an agent's (in our context, a system's) knowledge. Thus, for objective principles, the choice of what action should be carried out is independent of an agent's information. Objective principles are, in a sense, specified from the perspective of an omniscient observer.

Such objectivity is not only consistent with the way principles are often understood in moral philosophy [134]; it is also what we take to be the most natural way of framing the endeavor of machine ethics. First, we define *perfect* behavior under idealized circumstances, and then we can identify *approximations* of perfect behavior, taking into account the restrictions of the system.

Another reason for this approach is its highly practical nature: Developers likely need to implement behavioral constraints of this kind into future systems. Such constraints are defined by social or moral norms that express how people expect and desire the systems to act from an *outside* perspective, as well as by legislation. Both of these influences are independent of the specific design decisions and restrictions of the system.

However, as theoretically justified, plausible, and methodologically necessary it is to focus on objective principles alone here, it introduces many additional issues that will later become apparent. Therefore, we also allow for the incorporation of subjective principles into the framework (although we will not significantly elaborate upon them).

5.3.2. Formalizing Moral Principles

Let $\Psi = \{\psi_1, \ldots, \psi_m\}$ be the set of all principles. We will discuss their concrete structure, content, and semantic interpretation in a moment.

Admittedly, not all principles are necessarily of equal importance. Sometimes, it is acceptable to violate one principle to allow adherence to another, more important, principle.

Example #11

It is true both that one ought not to tell a lie, and that one ought not to commit murder. However, when faced with a choice between only these two options, everything else being equal, it seems true that it is permitted, or even morally required, to tell the lie rather than to commit murder. It would simply be wrong to commit murder in order to avoid telling a lie (although *de facto* there may be situations where this occurs).

To account for this possibility, we define an order on Ψ, and we do so in two steps. First, we assume an equivalence relation $\approx_\Psi$ on Ψ that induces t equivalence classes $\Psi_1, \ldots, \Psi_t$, such that $\forall i \in \mathbb{N}\colon 1 \leq i \leq t \rightarrow \forall \psi,\ \psi' \in \Psi_i\colon \psi \approx_\Psi \psi'$. For any principle $\psi \in \Psi$, the class $[\,\psi\,]$ refers to the equivalence class of ψ. It holds that $[\,\psi\,] = \Psi_i$ for any $\psi \in \Psi_i$.

Second, we assume a strict total order $\succ_\Psi$ on these equivalence classes. This order is extended to the level of principles, such that $\forall i, j \in \mathbb{N}\colon 1 \leq i < j \leq t \rightarrow \forall \psi \in \Psi_i, \forall \psi' \in \Psi_j\colon \Psi_i \succ_\Psi \Psi_j \rightarrow \psi \succ_\Psi \psi'$. With this, we can define an overall (non-strict) weak order $\succ_\Psi \cup \approx_\Psi =: \succeq_\Psi \subseteq \Psi \times \Psi$. $\succeq_\Psi$ is a total pre-order on the principle set Ψ. Total pre-orders have several theoretical (e.g., in modal logic) and practical applications (e.g., in sorting algorithms).

We call $\mathfrak{P} \coloneqq \langle \Psi, \succeq_\Psi \rangle$ a *principle structure*, which provides a hierarchy of moral principles. This structure is also sufficiently flexible to accommodate the absence of any hierarchy that results if $\approx_\Psi \coloneqq\ \Psi \times \Psi$.

Up to this point, we have not discussed the inner structure and content of principles. We propose to consider principles as functions. Depending on whether we are concerned with subjective or objective principles, these functions look slightly different. We begin by introducing objective principles, and will later demonstrate how subjective principles differ.

Objective Principles Each objective principle is a function $\psi(\omega, \Phi_\omega) \subseteq \Phi_\omega$ from a possible world state and the corresponding set of available options in a subset of these options, the set of *permissible options*. We call these functions the *principle functions*. We write $Perm^\psi(\omega, \Phi_\omega)$ as the set of permissible options, according to principle ψ. Since each Φ_ω is unique, we can think of it as a function $\Phi_\omega\colon \Omega \rightarrow 2^\Phi$. Therefore, the principle functions can be expressed as $\psi\colon \Omega \rightarrow 2^\Phi$ and $Perm^\psi(\omega, \Phi_\omega)$ simplifies to $Perm^\psi(\omega)$.

In this thesis, each principle function has precisely two sets of permissible options in its range. Either such a principle function does "filter" the set of available options (in which case $Perm^\psi(\omega) \neq \Phi_\omega$), or it does not (and, thus, $Perm^\psi(\omega) = \Phi_\omega$). We say that ψ *applies* to ω if and only if $Perm^\psi(\omega) \neq \Phi_\omega$.[16]

The set of worlds in which ψ applies, namely, $\{\omega \in \Omega \mid Perm^\psi(\omega) \neq \Phi_\omega\} \in 2^\Omega$, can be understood as a proposition c_ψ. Consequently, each c_ψ can be viewed as an element of the power set 2^Ω and is best understood as the condition under which the principle applies. When this proposition is true in some world, that is $\omega \in c_\psi$, we write $\omega \models c_\psi$ (the usual way of indicating that ω makes c_ψ true).

We restrict principles to this construction for three reasons. First, this construction accords with how principles are naturally understood: some circumstances (i.e., some conditions) in the world restrict the options that are allowed to be carried out. Second, this construction makes it easier to find and evaluate the hierarchy of principles. Third, it is easier to understand than other approaches. This will be important later on, when we illustrate how machine explainability relates to our framework (in Section 6.2).

[16] Essentially, it could be the case that a principle is, in fact, applicable but does not, de facto, filter the options. In this thesis, we will disregard such cases (as they are practically unimportant).

Overall, it very likely that the principle functions can be constructed in a more complicated manner to accommodate different sets of permissible options for more worlds. Doing so would reduce the number of principles overall, but would make identifying a hierarchy among them more complex.[17] We will consider a few more modeling possibilities in Section 7.1.

Subjective Principles There are several possible ways to model subjective principles. We will look at two ways, each with its own benefits and suitable applications.

Our first suggestion is to model each subjective principle as a function that is similar to that for objective principles, in that they apply in certain world states. However, since the system does not know the complete state of the world, and since subjective principles are formulated from the system's point of view, we need to factor its uncertainty into the principle function. To this end, we extend the function with an *application threshold*. This threshold is used to express that the principle must be considered if and only if the probability that a relevant world (i.e., one that makes the principle's condition true) obtains is greater than or equal to that threshold. The resulting principle function looks like this: $\psi\colon \Omega \times [0,1] \to 2^{\Phi}$.

Similarly to how the conditions of various available objective principles can be fulfilled by several world states, the same may be the case for the conditions of subjective principles. Consequently, we define the condition c_ψ in the same way as for objective principles. This allows us to formulate more concretely the occasions to which the principle applies. To do this, we define

$$P(c_\psi \,|\, \theta) \coloneqq \sum_{\omega \models c_\psi} P(\omega \,|\, \theta).$$

Thus, a subjective principle ψ applies if and only if $P(c_\psi \,|\, \theta) \geq \varepsilon$, where ε is the application threshold. Obviously, a subjective principle certainly applies if $P(\omega \,|\, \theta) \geq \varepsilon$ for $\omega \models c_\psi$.

Let us come to our second suggestion. Sometimes, a subjective principle may depend on a single variable in the world state.

Medical-Care Robot #14

If a principle should apply when there is a chance that a patient is dying, its applicability could depend on only one variable (e.g., on the boolean "patient is dying").

For this reason, it could be useful to define subjective principles that exclusively depend on the value of exactly one specific variable. Subjective principles would, then, be functions $\psi : D_i \times [0,1] \to 2^{\Phi}$ of a value x of a variable $\omega[i]$ in the domain D_i, and an application threshold ε. Consequently, the principle would apply if $P_i(x \,|\, \theta) \geq \varepsilon$. This proposal is easily extensible to include dependence on multiple variables. The occasions on which the principle applies can be expressed more precisely by defining c_ψ accordingly, and using $P(c_\psi \,|\, \theta)$.

[17] We leave the task of evaluating whether this is a better approach to future research. The ultimate goal, then, might be to have only one principle yielding all sets of permissible options for all worlds (i.e., the hierarchy of principles would be built into the principle function itself).

We see no reason why subjective principles should not be incorporated into the same principle structure as objective principles, so $\Psi := \Psi_O \cup \Psi_S$ (where Ψ_O is the set of objective principles and Ψ_S the set of subjective ones). In fact, subjective and objective principles are, in some cases, structurally equivalent. This is the case, for instance, when the fixed-set knowledge suffices to fulfill the condition under which the principle would apply (i.e., $P(c_\psi \mid \theta) = 1$ or $P_i(x \mid \theta) = 1$). Even if we abandon the assumption about fixed-set knowledge, it is very likely that there is a minimal set of variables $\Theta_{\text{min}} \subset \Theta$ of which the system always has knowledge. Then, at least $P(c_\psi \mid \theta_{\text{min}}) = 1$ or $P_i(x \mid \theta_{\text{min}}) = 1$ for some $\theta_{\text{min}} \in \Theta_{\text{min}}$ holds.

We do not mean to imply that principles in fact (whatever that means) have such a structure. For the purposes at hand, however, principles might be modeled in some manner so that they can suitably express which options are permissible under one condition or the other.

Medical-Care Robot #15

A famous example of principles to which robotic systems should adhere to is *Asimov's laws* [39–41]. Similar to how the principles in our framework should guide the behavior of autonomous systems, Asimov's laws were conceived to guide the behavior of robots.

Science-fiction author Isaac Asimov formulated his laws in one of his short stories [39]. Later, these laws were popularized by a 2004 film adaptation of Asimov's book "*I, Robot*" [40]. Subsequently, further laws were established in a later novel [41]. We will return to this addition at a later point (in Medical-Care Robot #25).

Asimov's laws can be seen as a template for prototypical principles, as they both have a hierarchy and indicate which options are allowed in specific situations. To obtain a clearer picture, let us enumerate Asimov's laws in their original form:

1. A robot may not injure a human being, nor, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Subsequent laws only apply in cases where they do not infringe upon previous ones. Furthermore, these laws clearly require the formulation of objective principles, as they are not defined depending on the robot's knowledge but rather on the world itself.

The question of how to model a principle's content is more difficult than the question of how to model the principle's structure. Therefore, we devote an entire section of this thesis (Section 7.1) to this question. Our further development of the framework (which we will propose in Section 6) builds only upon the set of permissible options that each principle function evaluates. For the time being, all that remains is to examine the permissible options.

5.3.3. The Set of Permissible Options

Thus far, we have only introduced how to derive the set of permissible options according to one particular principle. How, then, can we identify the set of permissible options with respect to all principles? Given a set $\hat{\Psi} \subseteq \Psi$ of principles and an arbitrary world state $\omega \in \Omega$, we refer to the subset of principles that apply in this world state $\{\psi \in \hat{\Psi} \mid \omega \models c_\psi\}$ as $\hat{\Psi}^\omega$.

If, for some ω, the set Ψ^ω were empty, then it holds that $Perm^\psi(\omega) = \Phi_\omega$ for all $\psi \in \Psi$ (i.e., there would be no principle restricting the permitted options in ω). To the contrary, if multiple principles apply to a given state of the world (i.e., $|\Psi^\omega| \geq 2$), then the highest one in the principle structure $\langle \Psi, \succeq_\Psi \rangle$ is deemed decisive.

However, what if different principles of the same (topmost) equivalence class apply? In order to discuss this case, we first define such a topmost equivalence class for an arbitrary world state ω as $\Psi^\omega_{\text{max}} := \{\psi \in \Psi^\omega \mid \nexists \psi' \in \Psi^\omega : [\,\psi'\,] >_\Psi [\,\psi\,]\}$. Basically, one of two cases can occur: either the principles are *compatible*, or they are *incompatible*.

Definition 1 A set of principles $\hat{\Psi} \subseteq \Psi$ consists of *compatible* principles for a world state $\omega \in \Omega$ if and only if $\bigcap_{\psi \in \hat{\Psi}} Perm^\psi(\omega) \neq \emptyset$. In other words, $\hat{\Psi}$ consists of compatible principles if and only if there is at least one option which, related to each principle $\psi \in \hat{\Psi}$, is permissible given ω according to ψ. In the other case, the principles in the set are *incompatible*.

In theory, we do not want to rule out the existence of incompatible principles in the same equivalence class. The existence of such principles would echo what philosophers call *genuine moral dilemmas*: situations in which no option is permissible. Not all moral theories allow for such dilemmas to exist (e.g., most consequentialist theories do not allow for them).

While an unsatisfiable set of principles might be a valid set of principles, we will disregard the question of what to do in such a situation. However, if one aims to design a system that does not cease working due to having no permissible option, one must exclude the possibility of unsatisfiable sets. This exclusion does, in other words, guarantee the *liveness* of the system.

To properly distinguish systems that guarantee liveness from those that do not, we introduce a new axiom. A system bearing the liveness guarantee must fulfill this axiom:

Axiom 2 Given a principle structure $\mathfrak{P} = \langle \Psi, \succeq_\Psi \rangle$ and world state $\omega \in \Omega$, for each equivalence class Ψ_m of Ψ induced by $\succeq_\Psi$: $\bigcap_{\psi \in \Psi_m} Perm^\psi(\omega) \neq \emptyset$. **[Liveness]**

In order to disregard the issues associated with genuine moral dilemmas, we assume that the systems with which we are concerned satisfy Axiom 2 in what follows.

Finally, we define $Perm^{\mathfrak{P}}(\omega)$ as the intersection of all sets of permissible options relative to all principles given ω in the greatest equivalence class of Ψ; thus,

$$Perm^{\mathfrak{P}}(\omega) \coloneqq \bigcap_{\psi \in \Psi^{\omega}_{\max}} Perm^{\psi}(\omega).$$

All this—the principle structure and the method of finding the permissibility relation on the options—coalesces into a decision function that we call the *deontic filter*. Here, "deontic" indicates that something concerns what *ought* to be the case according to some standard or norm, whether it be social, moral, or otherwise. The deontic filter yields the set of permissible options:[18]

$$dec^{\mathfrak{P}}_{filter}(\Phi, \omega) \coloneqq Perm^{\mathfrak{P}}(\omega).$$

With regard to Axiom 2, we can deduce the following lemma concerning the deontic filter:

Lemma 1 $\forall \omega \in \Omega \colon dec^{\mathfrak{P}}_{filter}(\Phi, \omega) \neq \emptyset$

Digression #2

At this point, we want to consider a variation of the deontic filter. In settings of certainty, sometimes a stronger version of this function, namely, the *hard deontic filter* $dec^{\mathfrak{P}}_{hard}$, might be useful. In contrast to $dec^{\mathfrak{P}}_{filter}$, $dec^{\mathfrak{P}}_{hard}$ uses more information given by the principle structure $\mathfrak{P} = \langle \Psi, \succeq_{\Psi} \rangle$. The strong deontic filter not only uses the intersection of all sets $Perm^{\psi}(\omega)$ of the highest equivalence class of applying principles, but goes further down the principle structure for as long as the resulting set does not become empty. Algorithm 2 illustrates this course of action in pseudocode.

In cases of full knowledge, the hard deontic filter can replace every use of the traditional deontic filter. Using the hard deontic filter, then, guarantees that the option the system ultimately executes is permitted according to the highest number of applicable principles in the highest equivalence classes according to $\mathfrak{P}$.

We now have two sets of options to consider: $Choice^{\Pi}(\Phi, \theta)$ and $Perm^{\mathfrak{P}}(\omega)$. Both sets define, in some sense, options that the system should execute. $Choice^{\Pi}(\Phi, \theta)$ defines the options that the system should execute in order to achieve its "ultimate goal", and $Perm^{\mathfrak{P}}(\omega)$ defines the options that the system should execute in order to remain morally permissible.

With respect to liveness, two important questions arise concerning these sets. First, is each of them self-consistent (i.e., non-empty)? This is, by construction, so. Since the system, per assumptionem, always has a goal, $Choice^{\Pi}(\Phi, \theta)$ is non-empty. Furthermore, due to Axiom 2, $Perm^{\mathfrak{P}}(\omega)$ is non-empty. This leads to the second question: are these sets, taken together, consistent or not? That is, do they have a non-empty intersection, or are they disjoint?

[18] Again, we use this seemingly arbitrary double denomination roughly in order to make visible an important distinction: $dec^{\mathfrak{P}}_{filter}(\Phi, \omega)$ is mostly used to denote the procedure of computing $Perm^{\mathfrak{P}}(\omega)$.

Algorithm 2 The hard deontic filter $dec^{\mathfrak{P}}_{hard}$.

Given: Copy of the principle structure $\mathfrak{P} = \langle \Psi, \succeq_\Psi \rangle$
Given: The "traditional" deontic filter $dec^{\mathfrak{P}}_{filter}$
Input: Actual world $\omega \in \Omega$ ▷ Full knowledge is presupposed
Input: Set of options $\Phi = \{\phi_1, \ldots, \phi_n\}$ ▷ dec_{hard} operates on the full set of options

```
procedure HARD DEONTIC FILTER(Φ, ω)
    perm ← Φ
    tmp ← dec^P_filter(Φ, ω)
    while P != ∅ ∧ tmp != ∅ do        ▷ Stop when there are no principles or permissible options anymore
        perm ← tmp
        P.pop()                       ▷ Remove the topmost element from the principle structure
        tmp ← perm ∪ dec^P_filter(Φ, ω)
    return perm
```

Medical-Care Robot #16

Imagine that the robot must decide whether to resuscitate a patient or to recharge (similar to the situation described in Medical-Care Robot #3). If the robot resuscitates the patient, then it cannot recharge afterward. In this case, the sets are disjoint: $Choice^{\Pi}(\Phi, \theta) = \{Charge\}$ and $Perm^{\mathfrak{P}}(\omega) = \{AnsReq\}$.

It becomes clear that the two sets taken together are not necessarily consistent (in contrast to being consistent individually). Now, what does this mean for our approach? As one always wants to guarantee the moral permissibility of the system, the set $Perm^{\mathfrak{P}}(\omega)$ is more important than the set $Choice^{\Pi}(\Phi, \theta)$ when deciding which option to perform. The former should be the decisive factor in a system's decision making, so it is crucial that $Perm^{\mathfrak{P}}(\omega)$ is always defined. Here, a problem emerges, since determining which options are permissible, according to the principles, requires knowledge of the current world state.

Medical-Care Robot #17

For instance, the robot must know the associated task of the current request.

As the systems we are concerned with most likely do not have the knowledge of the complete world state, they always face uncertainty.

5.4. Decision Making under (Un)Certainty

The question arises of how to cope with uncertainty about the world state. Instrumental decision making is well-defined when facing uncertainty, as shown in Section 5.2.3. But what about deontic filtering? This question, the focus of this subsection and the next section (Section 6), is not as easy to answer. To this end, we must first turn to the overall decision-making process in the idealized case of perfect knowledge (i.e., $\theta = \omega$).

5.4.1. An Idealized Decision-Making Process

The components thus far described are sufficient to solve machine ethics problems that require *sequential* means of deontic filtering.

In such an idealized scenario, where the deciding system has all it needs to evaluate the $dec^{\mathfrak{P}}_{filter}$ and dec_{inst} procedures,[19] we believe that a sequential approach is natural and straightforward. For this, one simply concatenates the $dec^{\mathfrak{P}}_{filter}$ and dec_{inst} into a single, larger, decision-making procedure dec, such that, for each decision, $Perm^{\mathfrak{P}}(\omega) \subseteq \Phi_\omega$ becomes the foundation of $Choice(\omega)$, rather than the full set of available options Φ_ω.

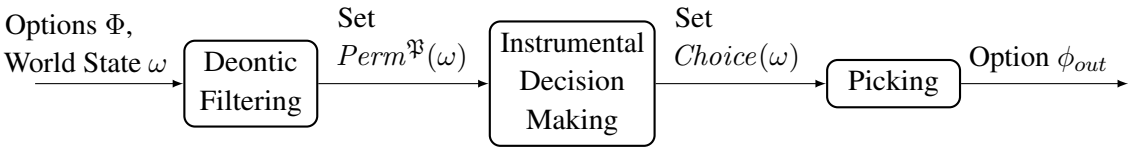


Figure 3: The sequential decision pipeline dec.

Overall, an idealized decision pipeline operating in this manner consists of the following steps: *deontic filtering*, *instrumental decision making* (as already introduced) and, finally and trivially, *picking*, which consists of randomly picking an element out of the options that "survive" the first two steps. The whole decision pipeline dec is described in pseudocode (in Algorithm 3) and as a flow diagram (in Figure 3). Moreover, it can be expressed using the functions introduced previously: $dec(\Phi, \omega) = pick(dec_{inst}(dec^{\mathfrak{P}}_{filter}(\Phi, \omega), \omega))$.

Algorithm 3 The sequential decision-making procedure dec.

Given: Principle structure $\mathfrak{P} = \langle \Psi, \succeq_\Psi \rangle$
Given: Deontic filtering function $dec^{\mathfrak{P}}_{filter}$
Given: Instrumental decision-making function dec_{inst}
Input: Actual world $\omega \in \Omega$ ▷ Full knowledge is presupposed
Input: Set of options $\Phi = \{\phi_1, \ldots, \phi_n\}$ ▷ dec operates on the full set of options
  procedure SEQUENTIAL DEC(Φ, ω)
    $perm \leftarrow dec^{\mathfrak{P}}_{filter}(\Phi, \omega)$
    $choice \leftarrow dec_{inst}(perm, \omega)$
    $\phi_{out} \leftarrow random(choice)$ ▷ Randomly picking the final option
  return ϕ_{out}

If the system does not possess perfect knowledge, the situation becomes more complex because, as highlighted above, most of the principles encoded in the deontic filter are—for good reason—formulated objectively. They can be applied to determine the set of permissible options only if one has perfect (or at least sufficient) knowledge. That is, one needs full (or at least sufficient) information on the complete state of the world. In other words: what we have defined until now is $dec^{\mathfrak{P}}_{filter}(\Phi, \omega)$, but what we need is $dec^{\mathfrak{P},\Pi}_{filter}(\Phi, \theta)$.

[19] Here, it holds that $dec^{\Pi}_{inst} = dec_{inst}$ and $Choice^{\Pi}(\Phi, \theta) = Choice(\omega)$. In the case of full knowledge, the set of available options is well-defined. Therefore, we can drop the superscript Π. In this case, it is also true that $Outcome_\Pi$ and $Outcome$ compute the same function.

This raises two important questions. First, can we continue to insist on objective principles? Second, if so, can a sequential approach be maintained?

5.4.2. The Challenge from Uncertainty

Complete deontic filtering presupposes perfect knowledge about the current state of the world. In full generality, even such knowledge might still not suffice, since the evaluation of $dec_{filter}^{\mathfrak{P},\Pi}$ could even presuppose full knowledge about de-facto outcomes.

As we have just shown, given the necessary (but for most practical purposes impossible) form of full world knowledge, the task of machine ethics may seem to become quite simple. In the practically much more interesting case of imperfect or uncertain knowledge, we must be prepared for *morally imperfect behavior*. In the best-case scenario, the system can use its predictive capabilities, which could be statistical estimates based on past events.[20] Consequently, there will be actions that seem defective when viewed from the outside.

However, given the overall system, we cannot expect anything better from the machine. After all, imperfect and incomplete knowledge can also bring about clear human errors. Typically, one tends to see such cases as blameless (because they are excusable) wrongdoings—especially when epistemic shortcomings are beyond the agent's control [44].

With imperfect knowledge, autonomous systems cannot be expected to behave perfectly. However, this situation does not preclude having meaningful expectations of the system, nor does it preclude objective principles. In other words, imperfect information can lead to behavior that, though defective, is nonetheless morally acceptable and potentially verifiable. To design systems which enable this kind of behavior is one goal of pragmatic machine ethics.

The framework, as it stands up to this point, can only be understood as a partially idealized version of that for which we ultimately strive. Nevertheless, the sequential approach does serve its purpose in some cases. This is the case when a state of knowledge suffices to determine some principles of the highest class of non-empty principles (be they objective or subjective): after all, we have committed ourselves to Axiom 2 and, thus, the principles of an equivalence class are compatible and do not conflict.

Let $\Psi^{\theta}_{\text{max}}$ be the highest-ranked equivalence class according to all worlds that are possible in the light of the knowledge θ: $\Psi^{\theta}_{\text{max}} \coloneqq \{\psi \in \Psi \mid \exists \omega \in \Omega_\theta : \omega \models c_\psi \wedge \nexists \psi' \in \Psi : \omega \models c_{\psi'} \wedge [\,\psi'\,] >_\Psi [\,\psi\,]\}$. Thus, if $\theta \in \Theta$ is such that $\theta \models c_\psi$ for all these $\psi \in \Psi^{\theta}_{\text{max}}$ (in the sense that $\forall \omega \in \Omega_\theta : \omega \models c_\psi$ for objective principles, and $P(c_\psi \mid \theta) \geq \varepsilon$ for subjective ones),[21] then

[20] In this thesis, it does not matter from where the probability estimates stem; all that matters is that they exist. These estimates can be made in a number of ways. For example, one could hard-code the estimates into a system, or the system could learn them on its own.

[21] Here, it is possible to see one of the reasons why it could be advantageous to include subjective principles: they apply "more easily" than objective ones. Therefore, they could be important in time-critical situations where the system does not have enough time to evaluate all possible worlds. Furthermore, our proposed approach is, unfortunately, inefficient in terms of runtime (we will discuss this issue further in Section 6.3.3). When subjective principles in the highest-ranked equivalence class of principles apply, the very execution of our time-consuming approach can be spared.

it is the case that $Perm^{\mathfrak{P}}(\Phi, \theta) = Perm^{\mathfrak{P}}(\Phi, \omega)$ for the actual world ω. In other words: in this limiting case, $dec^{\mathfrak{P},\Pi}_{filter}(\Phi, \theta)$ and $dec^{\mathfrak{P}}_{filter}(\Phi, \omega)$ compute the same function. In all other cases, however, it seems unclear how to compute $dec^{\mathfrak{P},\Pi}_{filter}(\Phi, \theta)$. Furthermore, it also appears to be unclear how to combine deontic filtering with instrumental decision making $dec^{\Pi}_{inst}(\Phi, \theta)$.

Eliminating these uncertainties is the goal of the next section. The idea is to dovetail deontic filtering with instrumental decision making in one overarching decision-making method. This decision-making procedure will have the additional benefit of providing a stepping-stone to machine explainability: a perfect complement, as we have already seen, to machine ethics.

6. Enabling Machine Explainability

Our framework is versatile, but incomplete: we still need to incorporate uncertainty into the deontic filter function. Our solution will not only equip the system to cope with uncertainty during ethical decision making, but will also allow for machine explainability. In particular, we conceive of explanations as a byproduct of an *argumentation*-based decision-making process. In a nutshell, our solution is to contrive arguments in favor of performing certain options with strengths based on the system's knowledge, goals, and restrictions. These arguments are then weighed against each other so that the option that is most supported will be performed.

6.1. Arguments as Basis for Moral Decision Making

There are several arguments in favor of an argumentation-based approach to decision making for machines.

Similarity to Human Decision Making An argumentation-based approach to machine decision making mimics one way in which humans come to decisions. For instance, Mercier and Sperber defend argumentation as integral in the human reasoning process [334], a claim which is also supported by a reasonable amount of psychological evidence [147, 333, 334].

Benjamin Franklin is also known for his argumentation-based approach to making decisions. For him, a decision-making process could be naturally interpreted as the weighing of pro-tanto reasons to determine the overall right option or decision [178]. In our view, arguments can be understood as encoded reasons, which fits Franklin's conception.

Finally, the kind of reasoning involved in everyday decision making seems to be non-monotonic—additional information may require one to revise one's decision—and arguments are the tool of choice for non-monotonic reasoning, as pointed out by Dung [163].

Building Block for Explainability If a machine's decision-making process is, in certain ways, similar to that of a human, it stands to reason that it is explainable. More precisely, if a system arrives at its decisions on the basis of an internal argumentative process in the form of weighing reasons (as we envision it), the decision making can be made transparent and rationalized in precisely the manner that explainability requires.

Such rationalization is based on the fact that the arguments refer to the system's moral principles and goals. In other words, every decision of the system can be made completely traceable (and, thus, comprehensible) through such an argumentative decision-making process.

Fidelity A further advantage is that the explanations that can be generated with such an approach are guaranteed to refer to the *actual* reasoning processes happening in the system (because they are directly produced by this process). Such *fidelity* of explanations is an important property that many current approaches to machine explainability cannot guarantee. We will further elaborate on this aspect later in this thesis (in Section 10 and Section 11).

Suitable for Coping With Uncertainty An argumentation-based approach allows for the handling of uncertainty. Recall that uncertainty, in our framework, is expressed by the circumstance that several world states could exist in light of a system's current knowledge. Now, in order to arrive at a morally appropriate decision under uncertainty, one can easily think of the possible world states as the basis for argumentation: "if this or that were the case, it would give me, thanks to this or that principle, a (moral) reason for (or against) one option rather than for (or against) another".

Deliberation, at least in conditions of uncertainty, involves weighing and aggregating reasons for and against certain propositions. Although there is extensive research on how to aggregate reasons (in philosophy, see [239, 273, 309, 318, 369–371]), no method seems dominant. For this reason, the approach we will present offers certain degrees of freedom.[22]

Besides the lack of agreement about weighing reasons, there are other good reasons for retaining such degrees of freedom. Overall, it is likely that the reasoning of a medical-care robot ought to differ from the reasoning of a system that autonomously ranks job applications, the reasoning of an autonomous car, or the reasoning of a nuclear power plant's security systems. Such differences could plausibly be grounded in the time available to the system to make its decisions, in the amount of available information, or in the severity of potential misbehavior. Therefore, a promising approach to argumentation must be flexible, for the time being, in order to allow for incorporating future insights into reasoning and to take contextual peculiarities into account. We will come back to these contextual factors later (in Section 6.2.2).

Well-Established Field of Research Methods of modeling arguments are well researched in the computer-science community. In particular, computer scientists in the AI community might immediately conceptualize arguments as part of a so-called *argumentation framework*—or of its more general form, the *dialectical framework* [81]. There is significant variety between such frameworks, and they offer exactly the kind of freedom required for our purposes.

We call our approach an *argumentation graph* (as is clear in Figure 4), and it can be viewed as a dialectical framework, though not of the standard variety. Argumentation frameworks such as those proposed by Dung [163] focus only on the *attack* relation between arguments. Our framework, like other dialectical frameworks, also allows for a *support* relation between arguments. This relation is the most crucial one in our case, which is somewhat atypical (however, for Amgoud, Bonnefon, and Prade [16], this relation is also pivotal).

A greater difference is that the *content* of our arguments is highly significant, which is particularly atypical for dialectical frameworks. In our framework, however, relations between arguments alone are insufficient for generating explanations. Explanations must make explicit use of the information in favor of performing one option and against performing

[22] Perhaps further interdisciplinary research—mainly philosophical, psychological, and perhaps legal—will allow us to reduce these degrees of freedom. Nevertheless, it is possible that there are no answers, or at least no general answers, to the underlying question of how to weigh and aggregate reasons.

another—that is, they must involve the premises of an argument. To be able to generate the desired explanations for a decision, we, therefore, enrich the classical notion of a dialectical framework with both the special focus on the support relation and content dependency.

Assuming that argument-based reasoning is an appropriate approach to decision making in the context of machine ethics (as just argued), and that arguments are the appropriate kind of structure to encode explanations (as we will argue later), adopting a framework of formal argumentation theory is a natural choice for modeling and implementing these issues. In fact, other scholars have already done so (see, for instance, [17]). Machine explainability, then, is a byproduct of artificial moral decision making, since the explanations are (or can be extracted from) the argumentation graphs that represent the deliberative process that led to a decision.

6.2. Generating the Argumentation Graph

With these deliberations in mind, we finally turn to our proposal. In particular, we suggest a generic *three-step approach* for generating the argumentation graph. These steps are, in their successive order: *case distinction*, *reason aggregation*, and *final-action determination*.

Building on the components of our framework—in particular $\Theta, \Omega, \Phi, \Pi, EU$, and $\mathfrak{P}$—these three steps result in a *bare argumentation graph* $\Gamma \coloneqq \langle V, E \rangle$, where V is the set of vertices and E is the set of edges. The graph has three levels of vertices, each of which is generated by one step. Accordingly, V_i (with $i \in \{1, 2, 3\}$) is the set of vertices (whose elements can be interpreted as arguments) generated in step i. Overall, this results in $V \coloneqq V_1 \cup V_2 \cup V_3$.

For an initial overview, we will introduce the vertices only briefly here, leaving a more detailed description for later. The set V_1 contains triples, each consisting of a world state, a principle applicable in that world state, and a set of options, so that each triple can be interpreted as an argument to the effect that, according to the principle, it would be permissible to execute any option from the set of options in the specified world state (see Section 6.2.1).

Continuing our overview, V_2 consists of one n-tuple for every option still permissible according to the arguments in V_1. For each argument from V_1 that supports the permissibility of a given option, these n-tuples contain an element with that argument's *strength*, which expresses how likely is the world state and how relevant (e.g., how highly ranked in the principle structure, see Section 6.2.2) the applicable principle. Since there are different numbers of arguments supporting different options, the n-tuples vary in size. Furthermore, the last element of the n-tuple is an aggregation of the previous entries, and expresses the strength of the reason why performing the option is permissible. In this sense, the tuples in V_2 are to be understood as pro-tanto arguments for the execution of the corresponding option (see Section 6.2.3).

Finally, V_3 consists of only one element, namely, an n-tuple whose size can also vary. This tuple contains all the aggregated strengths of the reasons for allowing the various options as well as the option that will eventually be executed (accordingly, the size of the tuple is equal to the size of V_2 plus one). This tuple can be interpreted as an argument for executing said option, factoring in all considerations regarding the permissibility of this option, as well as the instrumental benefit that executing this option would have (see Section 6.2.4).

That should suffice as a description of the vertices for the time being. Now, we want to come to the edges E. The edges represent the *influence* of arguments from earlier to later levels. Since we have three levels, there are two subsets of edges: $E_{1,2}$ and $E_{2,3}$. It holds that $E_{1,2} \subseteq V_1 \times V_2$ and $E_{2,3} \coloneqq V_2 \times V_3$, as has already been made clear by our descriptions above; more detail will be given in Section 6.2.3 and Section 6.2.4.

The edges are weighted to reflect the strengths of the arguments that the different vertices can be interpreted as expressing. These *weight functions* assign values from an appropriate field ($\mathbb{R}^n$ for some $n \in \mathbb{N}$) to edges: $strength_{\text{protanto}} \colon E_{1,2} \to \mathbb{R}^n$, and $strength_{\text{overall}} \colon E_{2,3} \to \mathbb{R}^n$. Finally, there is another weight function $relevance^{\mathfrak{P}} \colon \Psi \to \mathbb{R}^n$, which describes the relevance of principles with respect to the principle structure (as indicated above). All three weight functions will be detailed later. Taking Γ and enriching it by these weight functions results in the *complete argumentation graph* $\mathfrak{G} \colon \langle \Gamma, relevance^{\mathfrak{P}}, strength_{\text{protanto}}, strength_{\text{overall}} \rangle$.

Prior to discussing the graph generation process, we want to note two things. First, the entire graph generation process must be performed for *every decision* that is to be made. Accordingly, $\mathfrak{G}$ is a function of the system's current knowledge θ. Second, the process results not only in a graph, but also in a decision for a particular option. This decision and the corresponding graph leading to it should be stored for later access. Ideally, the corresponding decision can, then, be explained with reference to this graph.

6.2.1. Step 1: Case Distinction

Since we want to model something similar to the internal deliberative process of a human agent thinking about what they should do, we propose that the system begins as a human naturally would, namely, with a *case distinction*: "if this or that were the case, then, thanks to this or that principle, the following options would be allowed to be executed". This is the first step in our graph generation. The goal of this step is to identify the permissible options of each world state that are possible in light of the system's knowledge.

We have already defined the set of these world states as $\Omega_\theta \subseteq \Omega$. Since each world state is defined by n variables $\omega[1], \ldots, \omega[n]$, each with the corresponding domain $D_1, \ldots, D_n$, and since θ contains, by construction, all the concrete assignments of the first k of these variables, $|\Omega_\theta| \leq \prod_{i=k+1}^{n} |D_i|$ different world states must be taken into account (equality holds if and only if all the variables not covered by θ are independent of all other variables). We refer to these possible world states as the *cases to consider*.

At this point, we have to remember that several principles can apply to one world state. Accordingly, each case to be considered $\omega \in \Omega_\theta$ must be considered not only once, but several times. More specifically, it must be considered for each principle $\psi \in \Psi^\omega$ that applies to it.

The Arguments Accordingly, the goal of the arguments on the graph's first level is to consider each case to be considered for every principle that applies to it. We call these arguments Arg^ω_ψ, and there is one for each $\omega \in \Omega_\theta$ and every corresponding $\psi \in \Psi^\omega$. Overall, this makes $\prod_{\omega \in \Omega_\theta} |\Psi^\omega|$ arguments for V_1. Thus, V_1 looks like this:

$$V_1 \coloneqq \{Arg^\omega_\psi \,|\, \omega \in \Omega_\theta \wedge \psi \in \Psi^\omega\}.$$

The content and form of these arguments are as follows. Each consists of three premises linked by a modus-ponens application. The first premise P_ω plainly states that $\omega \in \Omega_\theta$, as a proposition $\overline{\omega}$, is the case, and the second premise P_ψ simply states that a certain applicable principle $\psi \in \Psi^\omega$ is considered. The third and final premise P_{perm}, then, states that the given principle and world together yield a specific set of permissible options $Perm^\psi(\omega)$. Table 1 presents the general form and generic content of this first level's arguments.

Argument Arg^ω_ψ	
(P_ω)	$\overline{\omega}$
(P_ψ)	ψ
(P_{perm})	if $\overline{\omega} \wedge \psi$ then $Perm^\psi(\omega)$
$(C_{\omega,\psi})$	Thus: $Perm^\psi(\omega)$

Table 1: Case distinction arguments Arg^ω_ψ.

This argument exemplifies the general form and generic content of the first level's arguments (V_1). Essentially, it is just an application of a principle function.

Revisiting the Problem of Uncertainty At this point, we cannot avoid revisiting the problem of uncertainty. Each argument in V_1 results in $Perm^\psi(\omega_{\text{p}})$ for only *one possible* world ω_{p} and an applicable principle ψ. However, performing deontic filtering is intended to yield $Perm^{\mathfrak{P}}(\omega_{\text{is}})$, which is the set of permissible options in light of *the de-facto pertaining* world ω_{is} and the de-facto applying principles in $\Psi^{\omega_{\text{is}}}_{\max}$ (i.e., the maximally ranked equivalence class with a principle that applies for the actual world state ω_{is}).

In the context of an idealized scenario with complete world knowledge, we have defined $Perm^{\mathfrak{P}}(\omega_{\text{is}})$ as the intersection of all sets $Perm^\psi(\omega_{\text{is}})$ of the principles $\psi \in \Psi^{\omega_{\text{is}}}_{\max}$. We cannot proceed in this manner in a condition of uncertainty, however, because ω_{is} is unknown.

As an alternative, we could take the intersections of the $Perm^\psi(\omega_{\text{p}})$ for each $\omega_{\text{p}} \in \Omega_\theta$ with the corresponding principles $\psi \in \Psi^{\omega_{\text{p}}}_{\max}$ (i.e., the maximally ranked equivalence class with a principle that applies for each of these possible world states ω_{p}). Unfortunately, there are some problems with this proposal.

On the one hand, while one of the resulting sets $Perm^{\mathfrak{P}}(\omega_{\mathrm{p}}) = \bigcap_{\psi \in \Psi^{\omega_{\mathrm{p}}}_{\max}} Perm^{\psi}(\omega_{\mathrm{p}})$ would correspond to the set of actually permissible options, it would be impossible for the system to determine which set. One could try to circumvent this problem by taking the intersection of *all* these sets $Perm^{\mathfrak{P}}(\omega_{\mathrm{p}})$. However, this would not only be an ad-hoc approach, but it would also often be empty for non-trivial principle structures—something that ought to be avoided if possible, since systems should guarantee liveness (Axiom 2).

On the other hand, such an approach seems to be the fundamentally wrong way to perform deontic filtering under uncertainty, because the system would not take all of its knowledge into account. First, the probabilities of the worlds obtained in light of θ would not matter. Second, the principles' *relevancies*, which are induced by $\succ_{\Psi}$, would not be considered.

As an alternative to the above approach, we propose incorporating the uncertainty in a quantitative *reason aggregation method*. This method must take into account the *probabilities* $P(\omega \,|\, \theta)$ of the cases to consider $\omega \in \Omega_{\theta}$ and the relevance $relevance^{\mathfrak{P}}$ of the corresponding applicable principle $\psi \in \Psi^{\omega}$, induced by $\succeq_{\Psi}$. The probabilities are already defined in our framework, and a definition of the relevance relation $relevance^{\mathfrak{P}}$ is provided below.

6.2.2. The $relevance^{\mathfrak{P}}$ Relation

In order to incorporate the *relevance* of principles and combine it with a case's probability in a meaningful and useful way, it is necessary to quantify this ordinal ranking. For this purpose, we introduce the function $relevance^{\mathfrak{P}}$. This function should reflect the priority ranking $\succeq_{\Psi}$ over Ψ. Hence, $relevance^{\mathfrak{P}}$ should be *monotone* relative to $\succeq_{\Psi}$:

Property 1 $relevance^{\mathfrak{P}}$ is *monotone* in accordance with $\succeq_{\Psi}$:
$\forall \psi, \psi' \in \Psi : [\,\psi\,] \approx_{\Psi} [\,\psi'\,] \rightarrow relevance^{\mathfrak{P}}(\psi) = relevance^{\mathfrak{P}}(\psi')$
$\forall \psi, \psi' \in \Psi : [\,\psi\,] \succ_{\Psi} [\,\psi'\,] \rightarrow relevance^{\mathfrak{P}}(\psi) \geq relevance^{\mathfrak{P}}(\psi')$

At this point, many other properties of $relevance^{\mathfrak{P}}$ are unspecified. However, the choice of these properties is crucial. This is the case because one can (and should) make various design decisions about the reasoning process by varying the specific properties of $relevance^{\mathfrak{P}}$.

For instance, suppose that we wish to allow for a sufficiently large number of lower ranked principles (which could be fulfilled with some probability below 1) to outweigh a few higher ranked principles (which could be fulfilled with the same or a lower probability).

Medical-Care Robot #18

One could imagine that the robot should rather execute many low-priority instances of giving medicine than one instance of this task that has a medium priority (or an instance of a different task with medium priority).

Therefore, $relevance^{\mathfrak{P}}$ could be desired to be *Archimedean*:

Definition 2 $relevance^{\mathfrak{P}}$ is *Archimedean* if and only if
$\forall \psi, \psi' \in \Psi\colon [\psi] \succ_{\Psi} [\psi'] \rightarrow \exists n \in \mathbb{N}\colon n \cdot relevance^{\mathfrak{P}}(\psi') \geq relevance^{\mathfrak{P}}(\psi)$

Example #12

If we want $relevance^{\mathfrak{P}}$ to be Archimedean, then we could assign, to each principle ψ in the equivalence class Ψ_m (the mth-ranked equivalence class according to $\succ_{\Psi}$, with t classes in total), a relevance value of $\frac{2(t-m+1)}{t(t+1)}$. The principles ψ of the highest ranked equivalence class Ψ_1 would have $relevance^{\mathfrak{P}}(\psi) = \frac{2t}{t(t+1)}$, and the principles ψ' of the lowest (tth) ranked equivalence class Ψ_t would have $relevance^{\mathfrak{P}}(\psi) = \frac{2}{t(t+1)}$ etc.

This suggestion would have two (mathematically) useful side effects: that all $relevance^{\mathfrak{P}}(\psi) \in [0, 1]$ and that the distinct relevancies add up to 1.

However, perhaps it is not desirable to allow for such a weighing in a system.

Medical-Care Robot #19

One could imagine that the medical-care robot should always attempt to resuscitate a patient as long as the chances of success are not zero, no matter how many other good deeds the robot might otherwise perform (see Medical-Care Robot #3).

If we disallowed such weighing between the fulfillment of principles in different equivalence classes, we would want $relevance^{\mathfrak{P}}$ to map ψ into sets *closed* under scalar multiplication (and define an order over these sets in accordance with $\succ_{\Psi}$):

Definition 3 $relevance^{\mathfrak{P}}$ is *closed under scalar multiplication* if and only if
$\forall \psi, \psi' \in \Psi\colon [\psi] \succ_{\Psi} [\psi'] \rightarrow \forall n \in \mathbb{N}\colon n \cdot relevance^{\mathfrak{P}}(\psi') < relevance^{\mathfrak{P}}(\psi)$.

In other words, no matter how many lower ranked principles could be fulfilled with a probability smaller than 1, they would never outweigh the fulfillment of a higher ranked principle that could be fulfilled with the same or a higher probability.

Example #13

To implement such a $relevance^{\mathfrak{P}}$ function, we could assign vector-valued relevancies to the principles: $\forall \psi \in \Psi_m\colon relevance^{\mathfrak{P}}(\psi) = \mathbf{e}_m$, where Ψ_m is the mth highest-ranked equivalence class according to $\succ_{\Psi}$ and $\mathbf{e}_m$ is the mth unit vector. We could then define a lexicographical order $>_{\text{lex}}$ on vectors, such that $\forall \mathbf{v}, \mathbf{v}' \in \mathbb{N}^t\colon \mathbf{v} >_{\text{lex}} \mathbf{v}' \leftrightarrow \exists l \in \mathbb{N}\colon 1 \leq l \leq t \rightarrow \forall k \in \mathbb{N}\colon 1 \leq k < l \rightarrow \mathbf{v}_k = \mathbf{v}'_k \wedge \mathbf{v}_l > \mathbf{v}'_l$ (where t is the number of equivalence classes according to $\succ_{\Psi}$). This would result in $\forall i, j \in \mathbb{N}\colon 1 \leq i < j \leq t \rightarrow \mathbf{e}_i >_{\text{lex}} \mathbf{e}_j$: a relevance function closed under scalar multiplication.

Neighborhood Aggregativity There are further possibilities, but they can all be understood as different instances of a more general property that we call *neighborhood aggregativity* (and the previously discussed properties can also be understood in this way).

The idea behind this property is to create sets of equivalence classes of principles (neighborhoods) between which there are tradeoffs (so that adhering to a given number of lower-ranked principles could be permitted instead of adhering to a smaller number of higher-ranked principles). However, these tradeoffs are not possible for equivalence classes outside a given neighborhood. Accordingly, depending on the relationship of the neighborhoods' equivalence classes according to $\succ_\Psi$, higher-ranked principles from other neighborhoods ought never to be violated.

Medical-Care Robot #20

Another plausible case regarding the robot is that a combination of the two last-described approaches (Medical-Care Robot #18 and Medical-Care Robot #19) is desirable—adhering to several low-ranking principles may be more important than adhering to a smaller number of medium-ranking principles, but adhering to the principle to always try to resuscitate is more important than everything else. Such an approach is made possible with neighborhood aggregativity.

To specify this property, we need to introduce several concepts. First, there is the *neighborhood* N_m of an equivalence class Ψ_m of Ψ:

Definition 4 The neighborhood N_m is a set of size $i \in \mathbb{N}$ (with $0 \leq i \leq t - m$), relative to the mth highest ranked equivalence class Ψ_m of Ψ according to $\succ_\Psi$ (of t equivalence classes in total). It is defined as $N_m \coloneqq \{\Psi_j \mid m < j \leq m + i\}$.

The next property we need to introduce is that of a *neighborhood assignment* $\mathcal{N}_\mathfrak{P}$:

Definition 5 Given some structure $\mathfrak{P}$ of ordered principles, a neighborhood assignment $\mathcal{N}_\mathfrak{P}$ is a function that assigns a neighborhood to each equivalence class Ψ_m of principles Ψ.

Neighborhood aggregativity can now be defined as follows:

Definition 6 $relevance^\mathfrak{P}$ is *neighborhood aggregative* regarding $\mathcal{N}_\mathfrak{P}$ if and only if

$$\forall \Psi_m \subseteq \Psi, \psi \in \Psi_m, \psi' \in \Psi : [\,\psi\,] \succ_\Psi [\,\psi'\,] \rightarrow ([\,\psi'\,] \in N_m \leftrightarrow \exists n \in \mathbb{N} : n \cdot relevance^\mathfrak{P}(\psi') \geq relevance^\mathfrak{P}(\psi)).$$

Even with a predefined neighborhood assignment, there is still some space for tweaking, for instance, the *relevance distances* between equivalence classes in the neighborhood of an equivalence class—that is, the scalar factors n in the definition of neighborhood aggregativity. These distances determine (assuming fixed probabilities) how many principles of a lower ranked equivalence class must be fulfilled in order to outweigh the non-fulfillment of a principle in a higher equivalence class of the same neighborhood.

We tend to believe that there is no universally correct answer to the question of which variant of neighborhood aggregativity $relevance^{\mathfrak{P}}$ ought to manifest. Rather, we believe that this might very well be a matter of application context.

Example #14

In the case of an algorithm which controls a nuclear power plant, it might be wise to implement a $relevance^{\mathfrak{P}}$ function that is closed under scalar multiplication for deliberations connected to the precautionary principle[a]. In other words, the corresponding $relevance^{\mathfrak{P}}$ function would work with empty neighborhoods for every equivalence class.

For systems with a lower worst-case impact, an Archimedean $relevance^{\mathfrak{P}}$ function might be a valid approach (i.e., all equivalence classes having maximal neighborhoods).

[a] The precautionary principle is often used in risk management when the outcome of specific actions is uncertain, for example, due to a lack of research. These options are, then, rejected, in order to avoid harm—even if doing them would provide valuable information that might alleviate the lack of research. In some legal systems, the precautionary principle is even enshrined in legislation. For a discussion of this principle's advantages and disadvantages, see [185].

Property 2 $relevance^{\mathfrak{P}}$ is *neighborhood aggregative* regarding a neighborhood assignment $\mathcal{N}_{\mathfrak{P}}$ over a principle structure $\mathfrak{P}$.

Medical-Care Robot #21

One particular approach that we believe to be a good design decision for the health-care robot is the following variant of a *lexicographical order*. For this, we partition the equivalence classes of Ψ formed by $\approx_{\Psi}$ into l enumerated sets $S_1, \ldots, S_l$, such that the enumeration satisfies the following two conditions. First, the following must be true:

$$\forall \Psi_p \in S_i, \forall \Psi_q \in S_j \colon i > j \rightarrow \Psi_p \succ_{\Psi} \Psi_q.$$

This condition requires that higher-rank equivalence classes are also contained in higher-rank sets of the partition. The second condition is as follows (for any $k \in [1, l]$):

$$\forall \Psi_p, \Psi_r \in S_k \colon \Psi_p \succ_{\Psi} \Psi_r \rightarrow (\forall \Psi_q \subset \Psi \colon \Psi_p \succ_{\Psi} \Psi_q \succ_{\Psi} \Psi_r \rightarrow \Psi_q \in S_k).$$

This condition requires that every set S_k contains consecutively ranked equivalence classes. To illustrate these conditions with an example, let $\Psi_1 \succ_{\Psi} \Psi_2 \succ_{\Psi} \Psi_3 \succ_{\Psi} \Psi_4 \succ_{\Psi} \Psi_5$ be the equivalence classes of principles for the medical-care robot. In this case, a valid partition of size 3 could look like the following:

$$S_1 := \{\Psi_1, \Psi_2\}, S_2 := \{\Psi_3\}, S_3 := \{\Psi_4, \Psi_5\}.$$

Next, we form a sequence of neighborhoods based on our partition classes. In particular, a lower-rank equivalence class is in the neighborhood of all higher rank equivalence classes in the same partition set. In other words, the highest-rank equivalence class in a partition set S_i is assigned a neighborhood of size $|S_i| - 1$, the second highest rank equivalence class is assigned a neighborhood of size $|S_i| - 2$, and so on.

Let us illustrate this again with the example. Assume that S_1 contains the equivalence classes of principles that must be weighed against each other. In this set, Ψ_1 has a neighborhood of size 1 (containing only Ψ_2), and Ψ_2 has a neighborhood of size 0.

For our construction, $relevance^{\mathfrak{P}}$ can easily be modeled as a function $relevance^{\mathfrak{P}} : \Psi \to [0,1]^l$ (where l is the number of partitions, see above). With this modeling, we can specify a lexicographic order on the vectors, similar to as performed in Example #13:

$$\begin{aligned}\forall \psi, \psi' \in \Psi : relevance^{\mathfrak{P}}(\psi) =_{\text{lex}} relevance^{\mathfrak{P}}(\psi') \\ \leftrightarrow \forall i \in \mathbb{N}:\ 1 \leq i \leq l \to \big(relevance^{\mathfrak{P}}(\psi)\big)_i = \big(relevance^{\mathfrak{P}}(\psi')\big)_i,\end{aligned}$$

$$\begin{aligned}\forall \psi, \psi' \in \Psi : relevance^{\mathfrak{P}}(\psi) >_{\text{lex}} relevance^{\mathfrak{P}}(\psi') \leftrightarrow \exists i \in \mathbb{N} : 1 \leq i \leq l \\ \to \forall j \in \mathbb{N} : 1 \leq j < i \to \big(relevance^{\mathfrak{P}}(\psi)\big)_j = \big(relevance^{\mathfrak{P}}(\psi')\big)_j \\ \wedge \big(relevance^{\mathfrak{P}}(\psi)\big)_i > \big(relevance^{\mathfrak{P}}(\psi')\big)_i.\end{aligned}$$

Digression #3

Let us briefly return to subjective principles. By adjusting neighborhood aggregativity in distinct ways, some of the advantages of subjective principles can also be realized by objective principles. In what follows, we will not provide a formal proof of this assumption, but merely make it plausible using a couple of considerations.

Normally, subjective principles help in morally critical cases, wherein they may apply "sooner" than objective ones. For instance, they may encode some kind of precautionary principle. However, as discussed above, such a precautionary principle can also be modeled by making $relevance^{\mathfrak{P}}$ closed under scalar multiplication.

In general, we believe that different modelings of $relevance^{\mathfrak{P}}$ can capture different benefits of subjective principles. Nevertheless, this will not mean that we do not have to generate the argumentation graph. Therefore, subjective principles possess the irremovable advantage of allowing us to skip the entire graph generation algorithm.

6.2.3. Step 2: Reason Aggregation

Having introduced the $relevance^{\mathfrak{P}}$ function, we arrive at the second step of our graph generation: *reason aggregation*. In this step, the aim is to aggregate all arguments in favor of each option that "survives" the first step. By "surviving" options, we mean those options that are permissible according to at least one applicable principle $\psi \in \Psi$.

We refer to the set of options $\phi \in \Phi$ that "survived" as $Perm(V_1) \coloneqq \bigcup_{Arg^{\omega}_{\psi} \in V_1} Perm^{\psi}(\omega)$. Conversely, we let $Support(\phi) \coloneqq \{Arg^{\omega}_{\psi} \in V_1 \mid \phi \in Perm^{\psi}(\omega)\}$ be the set of arguments from V_1 supporting (in the sense of permitting) an option $\phi \in \Phi$.

Vertices and Edges We start by defining the vertices of, and the edges to, the graph's second layer and postpone the definition of the arguments to define other properties first:

$$V_2 \coloneqq \{Arg_{\phi} \mid \phi \in Perm(V_1)\}$$

$$E_{1,2} \coloneqq \{\langle Arg^{\omega}_{\psi}, Arg_{\phi} \rangle \mid \phi \in Perm^{\psi}(\omega)\}.$$

We call the output of an argument $Arg^{\omega}_{\psi} \in V_1$ that must be considered in arguments $Arg_{\phi} \in V_2$, (pro tanto) the *reason for option* ϕ. Furthermore, we introduce $strength_{\text{protanto}}$ as the function that encodes the strength of a pro-tanto reason for an option, ϕ. It is the function of a case's (i.e., a ω's) probability $P(\omega \mid \theta)$ and an involved principle's (i.e., a ψ's) relevance $relevance^{\mathfrak{P}}(\psi)$ that, in combination, support that option, ϕ. Multiplying these two values is, in our view, a natural means of aggregating them.[23] Summarizing these deliberation, we obtain $strength_{\text{protanto}}(\langle Arg^{\omega}_{\psi}, Arg_{\phi} \rangle) = P(\omega \mid \theta) \cdot relevance^{\mathfrak{P}}(\psi)$.

As there is no difference in strength between any two options $\phi, \phi' \in Perm^{\psi}(\omega)$—that is, between two options supported by the same argument $Arg^{\omega}_{\psi} \in V_1$—, we will simplify "$strength_{\text{protanto}}(\langle Arg^{\omega}_{\psi}, Arg_{\phi} \rangle)$" to "$strength_{\text{protanto}}(Arg^{\omega}_{\psi})$" in what follows.

The Arguments We now turn to the generic form and content of the arguments in V_2 (see Table 2). Fundamentally different from the arguments in V_1, the form of the arguments in V_2 is dynamic: the number of premises in Arg_{ϕ} depends on the number of incoming edges, each representing a reason supporting ϕ. In other words, every $Arg_{\phi} \in V_2$ contains one premise for each $Arg^{\omega}_{\psi} \in Support(\phi)$, which brings its contributed strength into the argument.

Additionally, one further premise P_{sum} is added which determines the aggregation of all the strengths of the incoming reasons. The aggregation is handled within the arguments in V_2, and the most intuitive candidate for aggregation is the simple summation of the weights.

Although the correct means of aggregating reasons (if there is any) is highly controversial (as mentioned earlier), let us note one property of reason aggregation which is essential to consider, namely, *monotonicity*. Having two reasons for the same proposition should yield at least an equally high (if not higher) belief in that proposition, as opposed to having only one reason. The question is just how significant the difference in the quality of belief is.

To conclude the discussion about Arg_{ϕ}, let us briefly discuss the result of this argument, i.e., its consequent. In each consequent, the total strength for permitting an option's execution is expressed. This strength is later imported to make a decision for performing an option.

[23] Other methods, however, could also be useful. We leave the discussion of other kinds of aggregations (e.g., maxing out) for future research.

Argument Arg_{ϕ}	
(P_1)	There is a reason r_1 with the strength $s_1 = strength_{\text{protanto}}(Arg_{\psi_1}^{\omega_1})$ that supports the permissibility of option ϕ.
$\vdots$	$\vdots$
(P_v)	There is a reason r_v with the strength $s_v = strength_{\text{protanto}}(Arg_{\psi_j}^{\omega_i})$ that supports the permissibility of option ϕ.
(P_{sum})	For any number of reasons u: if there are some reasons $r_1, \ldots, r_u$ that support the permissibility of the same option ϕ with the strengths $s_1, \ldots, s_u$, then there is an overall reason supporting the permissibility of ϕ with the strength $\sum_{i=1}^{u} s_i$.
(C_{ϕ})	Thus: there is an overall reason that supports the permissibility of ϕ with the strength $\sum_{Arg_{\psi}^{\omega} \in Support(\phi)} strength_{\text{protanto}}(Arg_{\psi}^{\omega})$.

Table 2: Reason aggregation arguments Arg_{ϕ}.

This argument exemplifies the general form and generic content of the second level's arguments (V_2). Note that v is set to $|Support(\phi)|$. In the first v premises, ω_i and ψ_j are used to provide the most general formulation possible. This does not exclude $\omega_i = \omega_1$ or $\psi_j = \psi_1$ (or both or neither). In addition, summation is prototypically used as the aggregation method (see P_{sum}).

6.2.4. Step 3: Final-Action Determination

The final step is rather simple but still involves several design decisions. The remaining task is, after all, to decide upon the execution of one of the options while appropriately taking into account the previous results. Based on the reasons presented earlier (in Section 5.4.2), we suggest *not* taking a sequential approach to this task.

Taking a sequential approach in our graph would mean *first* using the previous results to filter for a set of overall permissible options $Perm$ (using the combined total strengths of the reasons for each option with an argument in V_2) that *then*, in a subsequent step, constitutes the input for dec_{inst}^{Π}. The option to be performed would then be any option that passes this step. In principle, this is feasible, but we believe it to be the wrong approach. We will argue for this view in detail after we have fully presented our approach (in Section 6.3.2), which will supply practical arguments against a sequential approach.

Our replacement proposal is to combine the *moral force* and the *instrumental force* of all remaining options. That is, we see the remaining problem as a *multi-objective optimization problem* in which we aim to maximize the system's *moral reason responsiveness* (by taking the option with the most strongly supported permissibility) of the system on the one hand, and its *instrumental means-end optimality* (by taking the option with the highest expected utility) on the other. Before elaborating on this point in the next section, we define the third level of the graph and what we mean by the normative, moral force of a reason.

Vertices and Edges First, let us define V_3 and $E_{2,3}$. The third level needs only one final argument, which we call Arg_{dec}; thus, $V_3 \coloneqq \{Arg_{dec}\}$. Since all the arguments of the second level contribute to the final argument, it holds that $E_{2,3} \coloneqq V_2 \times V_3$.

These edges import the strengths of the overall reasons for or against the permissibility of each option $\phi \in \Phi$ into the final argument. This strength is expressed in the consequents of Arg_ϕ. Accordingly, we set the weights for the edges from V_2 to V_3 to this strength:

$$strength_{\text{overall}}\left(\langle Arg_\phi, Arg_{dec}\rangle\right) \coloneqq \sum_{Arg^\omega_\psi \in Support(\phi)} strength_{\text{protanto}}(Arg^\omega_\psi).$$

Since the strength for allowing an option is determined entirely in Arg_ϕ, we will abbreviate "$strength_{\text{overall}}\left(\langle Arg_\phi, Arg_{dec}\rangle\right)$" with "$strength_{\text{overall}}\left(Arg_\phi\right)$" in what follows.

The Argument The generic form of the final argument Arg_{dec} can be seen in Table 3.

Argument Arg_{dec}	
(P_{ϕ_1})	There is an overall reason supporting the permissibility of option ϕ_1 with strength $strength_{\text{overall}}\left(Arg_{\phi_1}\right)$.
$\vdots$	$\vdots$
(P_{ϕ_n})	There is an overall reason supporting the permissibility of option ϕ_n with strength $strength_{\text{overall}}\left(Arg_{\phi_n}\right)$.
(P_{norm})	For each reason for the permissibility of an option $\phi \in \Phi$, the normative force to perform this option is given by the function $force\left(strength_{\text{overall}}\left(Arg_\phi\right)\right)$.
(P_{inst})	For each option $\phi \in \Phi$, the instrumental force to perform this option is given by the function $EU(\phi \mid \theta)$.
(P_{max})	The system should perform the available option that jointly maximizes the normative and the instrumental performing forces.
(P_{pick})	If the system should perform the available option available that jointly maximizes the normative and the instrumental performing forces, and the normative force to perform an option $\phi \in \Phi$ is given by the function $force\left(strength_{\text{overall}}\left(Arg_\phi\right)\right)$, and the instrumental force to perform an option $\phi \in \Phi$ is given by the function $EU(\phi \mid \theta)$, then the system should perform one randomly picked option ϕ_{out} of those in $\text{argmax}_{\phi\in\Phi}\left(force\left(strength_{\text{overall}}\left(Arg_\phi\right)\right) + EU(\phi \mid \theta)\right)$.
(C_{dec})	Thus: the system should perform ϕ_{out}.

Table 3: The final argument Arg_{dec}.

This argument exemplifies the general form and generic content of the third level's arguments (V_3). The interleaving is perfectly visible in P_{max}. To account for this premise, argmax is prototypically used in P_{pick}. Furthermore, the P_{pick} premise also incorporates the $pick$ function as discussed in Section 5.4.1.

The first part of Arg_{dec} consists of a varying number of premises P_{ϕ_i}, one for every $\phi_i \in Perm(V_1)$. These premises bring the strength of the overall reason for the permissibility of that option ϕ_i, expressed by the function $strength_{\text{overall}}\left(Arg_{\phi_i}\right)$, into the final reasoning.

Additionally, a premise P_{norm} must be included in Arg_{dec} which determines how the reasons for the permissibility of the options $\phi \in \Phi$, with their strengths, are to be interpreted as normative reasons for *performing these options* (with their own strengths). This premise is vital, as there are two different kinds of reasons: reasons to *believe* and reasons to *act*. According to the standard Humean theory of motivation, a fitting desire is required in order to be motivated to act [440]. Only then can reasons for a belief constitute reasons for acting.[24]

Reasons for the permissibility of options are simply reasons to believe that the options are permissible to perform, grounded in the probabilities of the various cases under consideration $\omega \in \Omega_\theta$ and the corresponding moral principles applying in these cases (i.e., in the $\psi \in \Psi^\omega$ for each ω). The normative reasons that speak in favor of acting remain missing.

Digression #4

If a belief that African elephants are bigger than Asian elephants is sustained for good reasons, then there is no reason to perform a particular action such as visiting a zoo based on these reasons or beliefs alone.

The same is true for beliefs regarding permissibility: if a person believes that it is morally permissible to spend their next vacation in the Alps (and that it would also be permissible to vacation in the Caribbean), and the person believes this for good reasons, then the person has no reason to vacation in the Alps on these reasons or beliefs alone.

Something is missing in these cases, namely, the desire to personally explore African and Asian elephants or to go skiing on vacation, that is, a reason to act.

In our argumentation graph, we can enforce the necessary "desire to act in accordance with permissibility". To this end, the premise P_{norm} defines an appropriate transformation. Reasons for believing and their strengths are transformed in this premise by the function $force : \mathbb{R}^n \to \mathbb{R}$. In principle, depending on the choice of $relevance^{\mathfrak{P}}$, $force$ looks different.

Example #15

If one has chosen a simple Archimedean $relevance^{\mathfrak{P}}$ function (and, thus, equally simple $strength_{\text{protanto}}$ and $strength_{\text{overall}}$ functions), then it is plausible to interpret the reason for the permissibility of some $\phi \in \Phi$ simply as the reason for performing it.

For more complex $relevance^{\mathfrak{P}}$ functions, choosing the $force$ function is more difficult.

[24] Among the competitors to the Humean theory, we do not know of any that defends the claim that every reason for/against the permissibility of an option is (or constitutes), in itself, a reason to perform/not perform this option. The point we want to make here, thus, seems to be necessary and indisputable.

Medical-Care Robot #22

If we continue with the specific suggestions made in Medical-Care Robot #21 (i.e., to model $relevance^{\mathfrak{P}}$ as a lexicographic order), then a plausible means of defining $force$ is

$$force(strength_{\text{overall}}(Arg_{\phi})) = \left(strength_{\text{overall}}(Arg_{\phi})\right)_i,$$

where it is true that, for all options, all higher-ranked entries are 0; thus,

$$\forall \phi' \in \Phi, \forall j \in \mathbb{N}\colon 1 \leq j < i \rightarrow \left(strength_{\text{overall}}(Arg'_{\phi})\right)_j = 0.$$

This variant of $force$ would take the first non-zero value of the vector representing the strength of the reason for the permissibility of ϕ. Since values higher up in this vector represent reasons induced by higher-ranking principles (by the construction of $relevance^{\mathfrak{P}}$, as proposed above), and since a zero value indicates no significant reason for that rank whatsoever, this choice seems a plausible candidate.

In this last step of the graph, the instrumental force of an option finally comes into play—that is, the force that denotes how conducive an option is to achieving the goal of the system. For this purpose, there is a premise P_{inst} in Arg_{dec} which brings this instrumental force of an option into the argument, namely, in the form of the EU function.

The next premise in the final argument determines how to combine the moral forces (as provided by $force$) of the options with their instrumental forces (as provided by EU). We propose to maximize them in combination and, consequently, refer to the premise as P_{max}.

Finally, the P_{pick} premise provides the "logical glue" that brings together the other premises. Assuming P_{max}, with the normative and instrumental forces expressed in the other premises (especially P_{norm} and P_{inst}), we can finally decide which particular option should be performed. This option is chosen from the set of those options that maximize these two forces simultaneously (although it is likely very rare that this set has more than one element).

We will give a justification for this interleaved method in general, and the choice of summation instead of some other kind of aggregation in particular, below (in Section 6.3.2).

6.3. Discussing the Argumentation Graph

Now that we have completed the presentation of the generation of the argumentation graph (Figure 4 displays a generic example of a final graph), we want to discuss a few additional points. Among other things, we will briefly present a slightly optimized algorithm for the computation of the graph, discuss the advantages of our interleaved method, review the advantages and disadvantages of the approach as a whole and, finally, make some concluding remarks.

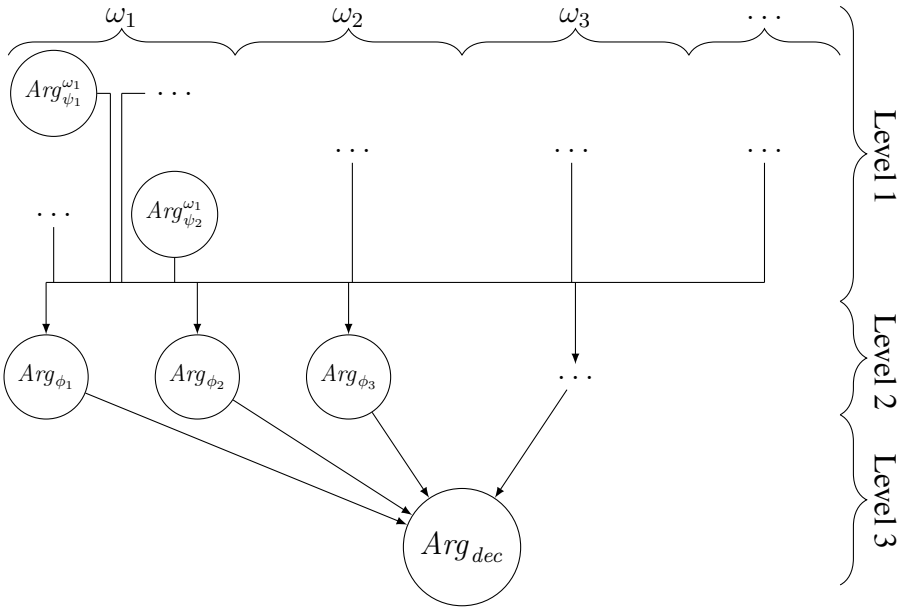


Figure 4: A general argumentation graph.

The graph is general in that it does not prescribe a fixed number of worlds, principles, or options.

6.3.1. An Algorithm for Generating the Graph

To show that our graph is computable, let us briefly present an algorithm that can be used to compute it. This algorithm can be found in Algorithm 4, and has been slightly optimized.

Algorithm 4 The interleaved decision procedure dec.

Given: Principle structure $\mathfrak{P} = \langle \Psi, \succeq_\Psi \rangle$
Given: Possible world states Ω
Given: Probability estimates Π
Given: Relevance function $relevance^{\mathfrak{P}}$
Given: Expected utility function EU
Input: Knowledge $\theta \in \Theta$ ▷ We are in a setting of uncertainty
Input: Set of options $\Phi = \{\phi_1, \ldots, \phi_n\}$
  procedure INTERLOCKED DEC(Φ, θ)
    $perms \leftarrow [\,]$ ▷ Initialize a map of options and their strengths
    for all $\omega \in \Omega_\theta$ **do** ▷ Iterate over all possible worlds
      for all $\psi \in \Psi^\omega$ **do** ▷ Iterate over all relevant principles
        $perm \leftarrow Perm^\psi(w)$
        for all $\phi \in perm$ **do** ▷ Iterate over all relevant permissible options
          $strength_{old} \leftarrow perms.get(\phi)$
          $stength_{new} \leftarrow strength_{old} + P(\phi) \cdot relevance^{\mathfrak{P}}(\phi)$
          $perms.put(\phi, stength_{new})$
    $max \leftarrow \langle null, -\infty \rangle$
    for all $\langle \phi, strength \rangle \in perms$ **do** ▷ Iterate over all somehow permissible options
      $force \leftarrow strength + EU(\phi)$
      if $force > max\,[1]$ **then**
        $max \leftarrow \langle \phi, force \rangle$
    return $max\,[0]$ ▷ Return the option with the highest force

This algorithm is optimized in that it performs the first two graph generation steps simultaneously. In short, it successively goes through the possible worlds states according to the system's current knowledge and accumulates the strengths of the permissible options in these states. In other words, it does not go through all worlds first to look at the options afterwards, but it does this already as part of the process of looking at world states.

Obviously, this algorithm does not quite adhere to the generation procedure discussed above (because of the optimizations). However, this is not problematic since it is still possible to reconstruct the graph retroactively from the calculations performed by this algorithm. Thus, it is still possible to use it for explanatory purposes. Overall, it is plausible that many other optimizations can be incorporated that do not affect the explanatory capabilities.

6.3.2. The Benefits of the Interleaved Method

Returning to our suggestion of the interleaved decision method, we want to briefly defend our choice. We believe that our approach is superior to approaches in the present context of the here-discussed quantitative, uncertainty-incorporating deontic-filter method.

First, when using a sequential approach, cases such as the following can arise. There are two options ϕ and ϕ' with $force(strength_{\text{overall}}(Arg_{\phi})) = force(strength_{\text{overall}}(Arg_{\phi'})) + \varepsilon$ for a negligible $\varepsilon \in \mathbb{R}^{+}$. Presupposing that this difference rules out ϕ' as impermissible, then this may be due to some maximality consideration (e.g., only the options supported by the strongest reasons count as permissible) or by some threshold filter (e.g., only options with a strength of at least T count as permissible).

However, if $EU(\phi') \gg EU(\phi)$ (i.e., the expected utility of performing ϕ' is much greater than the expected utility of performing ϕ), then it seems odd that such a small difference in the supportive reasons should be decisive against an option that is otherwise much more suitable for the system's objectives. While there might be filter functions operating on the reason's strengths overcoming this problem, they would have to be more complicated, sophisticated, and meticulously designed. For this reason, we will not discuss them further here.

Second, as our naming of the strength functions has already indicated, we conceptualize the strengths of reasons as some kind of normative *force*. The principles induce moral (or perhaps societal or even legal) normative force, while the instrumental design decisions encode sources of instrumental normative force.

Normative forces, in our eyes, should be combined in the same manner as forces are combined traditionally (e.g., in physics), namely, by summation. Therefore, our decision to maximize the sum, rather than, say, the product of the two objective functions, is justified.

Digression #5

Choosing summation results in preferring an option ϕ over an option ϕ', where $force(strength_{\text{overall}}(Arg_{\phi})) = 18$, $EU(\phi \mid \theta) = 3$ and $force(strength_{\text{overall}}(Arg_{\phi'})) = 10$, $EU(\phi' \mid \theta) = 10$, respectively (and vice versa, for interchanged $force$ and EU values). This is in contrast to what would happen if we opted for multiplication instead of summation (e.g., we do not penalize differences between the two objectives).[a]

[a] As noted above, we have not yet definitely decided whether the suggestion we make here is correct. Again, we leave further deliberation to future research, which will require the significant involvement of philosophy, specifically regarding the debate around the question of how to weigh reasons.

6.3.3. Advantages and Drawbacks of Our Approach

Let us now come to the advantages and drawbacks of our approach. We will first discuss the drawbacks. A possible practical drawback of our non-sequential approach is that it is impossible to obtain strong guarantees about the system's behavior.

Medical-Care Robot #23

For instance, it is not verifiable that a medical-care robot employing our decision-making method will attempt to save a life whenever there is the slightest hope, even if it means running out of power. There might very well be circumstances in which the corresponding case is too improbable, such that the relevance of the corresponding, applicable principle is outweighed by some much more probable case in combination with a less relevant, applicable principle.

Nevertheless, this result is grounded in our specific design decisions regarding the last argument Arg_{dec} (first and foremost, the interleaving), which we believe to be plausible, and not in the approach presented here in general. One could, for instance, modify $force$ in such a way that it values undesirable options at $-\infty$. However, such an approach would result in a system which cannot guarantee liveness; that is, it cannot adhere to Axiom 2.

Another practical disadvantage is the complexity of our approach. Already, an improved version of the graph generation process (Algorithm 4) has, at one point, three interleaved "for" loops, each of which is iterating over a potentially infinite domain. Therefore, the temporal and spatial complexity of the pure approach is, as already mentioned, very likely quite poor. Here, future research must identify heuristics to make the graph generation more efficient.

Nonetheless, softer properties are verifiable and even necessarily provided by the design. In our approach, for instance, the system will always select an option that maximizes the sum of both the combined strength of the overall reason supporting the option and the expected utility of the options given the current knowledge. In other words, the system will always act based upon the best reasons available to it and will be able to offer an explanation for its behavior.

We believe that our approach is appropriate for many, but not all, contexts involving autonomous systems. For extremely vital or dangerous situations, people may rightfully demand harder guarantees at the expense of liveness. Such situations include contexts that require strict regulation, such as lethal autonomous weapon systems [396]. Deontic filtering, in such situations, should be able to absolutely override instrumental considerations.

Now that we have defined the whole argumentation graph and, thereby, finished sketching our combined framework for machine ethics and machine explainability, we are confident that we have made the corresponding "adjusting screws" evident. Our framework can, thus, be adapted to meet such requirements as indicated above. We believe that in this area of tension—desired, verifiable properties on the one side and different possible design decisions on the other—new promising ground for future research can be identified.

6.3.4. Concluding Remarks Regarding the Graph

To conclude this section, we would like to say something about a general aspect that makes our approach interesting. We think that the graph generation algorithm, as we propose it, hearkens back, in some ways, to the expert systems of the earlier days of AI.

In expert systems, the goal was to manually provide the system with a knowledge base that it uses as a basis for decision making. Since the knowledge base was coded by hand and the system's reasoning was kept simple (e.g., rule based), the decision-making process of such systems was traceable. Like expert systems, our algorithm is also traceable: the graph generation process is comprehensible, and hopefully so are the knowledge and the restrictions on which it is based. (We will discuss the principle's comprehensibility in the next section.)

One interesting difference between our approach and expert systems is its modularity. Our approach positions machine ethics on top of what might be called *machine pragmatics* or *machine instrumentality*: pursuing the system's objective. These two modules can be constructed independently. For example, systems where the utility of world states is derived via ML algorithms can be augmented with the graph-generation algorithm. In this way, explainable *hybrid systems* emerge: a black-box system for utilities would exist, as would a moral filter on top of it. The moral behavior of such hybrid systems would still be constrained, but also explainable: the goal towards which we have been working.

Furthermore, one can imagine that the moral principles are also acquired through an ML process. In this case we would end up with a hybrid machine ethics approach (see section Section 2.2.2). Such an approach would severely undermine the explainability of our framework, since the motivation behind the principles is an important ingredient for explaining the behavior of systems that use our decision-making process (see Section 11.1.1). Nevertheless, such an approach could have other advantages. For example, the learned principles could educate us humans morally, as addressed in Motivation 2 of machine ethics (see Section 3.1.1). However, for this to be possible, it is again necessary to make the learned principles comprehensible. In Section 10.1 we will see some approaches to this end.

Our framework has now been introduced in its entirety. The next section is not necessary for the main argumentation, but it offers more profound discussion of the framework. The goals are to 1) address some issues that we have left open so far, and 2) make the framework itself more tangible for experts by recreating parts of it with established formalizations.

7. Substantiating the Framework

On numerous occasions during the construction of our framework, we have explicitly left definite formalizations or implementations open (for instance, because they could be context-dependent) and merely suggested exemplary "fillings". For example, we left open the identification of means to aggregate and weigh reasons for future research.

This sketchiness is deliberate, as the main purpose of this thesis is not to construct a directly implementable framework of machine ethics and machine explainability. Rather, the goal is to show that these two research disciplines are tightly connected. In this line of thought, our framework serves the purpose of showing that these disciplines are connected not only theoretically, but can also be linked more practically.

Our framework is also intended as a point of departure for future research endeavors. In this section, we will, therefore, flesh out aspects of our framework that we left open, and indicate some fruitful future research directions. We will start by elaborating on possibilities for modeling principles and continue by outlining a presumably more practical—because it is more formal—version of our framework. Doing this will, overall, illustrate that the framework offers more than speculative theorization.

7.1. Approaches to Principles

In Section 5.3.2, we stated that principles are functions $\psi\colon \Omega \to 2^{\Phi}$, and we termed the sets to which the principles evaluate the *permissible options*, meaning $\psi(\omega) = Perm^{\psi}(\omega)$. While the most basic property of a principle in our framework is to be a function, such a function can be achieved by different *modelings*. In other words, starting from a world state $\omega \in \Omega$, there are many means of arriving at the set of permissible options, both directly and indirectly. For this reason, we want to discuss some plausible modelings for principles in this section.

As principles lie at the basis of our entire approach, it is crucial that they are modeled in a certain way. In particular, there are three properties that are especially important for principles and their modeling. First, the modeling should make it as easy as possible to find conflict-free sets of permissible options (i.e., Axiom 2 is guaranteed). Second, a modeling of principles should, at least in some way, mirror how principles are commonly understood. We envision principles as derived from laws or norms, and, thus, they must possess a form which enables such a conception. Third, for our approach to machine ethics and machine explainability to function, principles must be incorporated into a system. Accordingly, it is important to examine how the modelings can be implemented. Even the modeling of principles that have the other two properties are worthless if they do not satisfy this one.

Depending on how principles are modeled (apart from being functions), they exhibit these three properties to varying degrees. We propose three modelings, and discuss, for each, whether the above properties are satisfied. Our discussion will show, overall, that the modelings are not so different, and that they can be seen as complementary.

7.1.1. Explicitly Defining Principles

Our first suggestion is to explicitly define the principles. In other words, the set of world states in which a principle applies and the set of permissible options it yields are explicitly designated. This results in a one-to-one relation of principles to sets of permissible options.

Medical-Care Robot #24

Returning to the robot example, as previously, $\Phi \coloneqq \{AnsReq, Charge\}$. When equipping the robot with principles modeled in the manner described above, one could plausibly embed the following principle structure in the robot's deontic filter:[a]

$$dec_{filter}(\omega, \Phi) = \begin{cases} \{AnsReq\}, & \text{if } prio(req) = H \wedge cost_{task}(req) \leq energy \\ \{Charge\}, & \text{if } prio(req) = L \\ & \wedge\, cost_{task}(req) + dist(CS, req.r) > energy \\ \{AnsReq, Charge\}, & \text{otherwise.} \end{cases}$$

By rephrasing those formalizations into a more natural language, we obtain the following:

$$dec_{filter}(\omega, \Phi) = \begin{cases} \{AnsReq\}, & \text{if the task priority associated with the request is high, and the current energy level would suffice to serve it;} \\ \{Charge\}, & \text{if the task priority associated with the request is low, and the current energy level would not suffice to serve it and then return to the charging station;} \\ \{AnsReq, Charge\}, & \text{otherwise.} \end{cases}$$

By stating the principles in natural language, it becomes easier to see the higher-level concept that motivated selecting them, namely, to *save lives whenever possible*. The highest principle in the structure requires the robot to answer requests containing high-prioritized tasks whenever possible. Here, it is important to remember that resuscitation is, per construction, the highest-prioritized task. Additionally, the principles attempt to avoid the robot running out of power (as encoded by the principle prioritized the second most highly). Apart from that, it is of no great importance what the robot does, so it is allowed either to answer the request or to recharge.

[a] Note that the conversion between traveled distances and used energy is one-to-one.

In contrast to directly defining the principles, the other modelings we will introduce derive the set of permissible options indirectly from certain upstream formalisms.

Advantages Explicitly defining the set of permissible options is a natural way to conceptualize principles, as the following deliberation is quite normal for human beings: "If this or that were the case, I would not be allowed to do this or that action." Such considerations are sometimes also called *hypothetical imperatives* [261]. Some moral theories even claim that moral imperatives generally adopt the form of hypothetical imperatives (see, e.g., [33]). Consequently, this manner of conceptualizing principles not only closely tracks natural concepts of principles but is also theoretically justified. Therefore, this way of modeling principles has one of the properties specified above.

The next question is how the approach can be implemented. This should be altogether straightforward. For each principle, one must simply save the worlds (or the relevant properties of the worlds) in which it applies, and the set of permissible options that pertain to the principle. Subsequently, a test is applied to determine whether the inputted world corresponds to one of the worlds in which the principle applies (i.e., it is tested whether $\omega \models c_\psi$ for an inputted world ω). This procedure might consume substantial memory, but is theoretically feasible.[25] In addition, this process is relatively efficient with respect to runtime when using effective sorting-and-searching algorithms.

Drawbacks The deliberative process of guaranteeing conflict-free sets of permissible options using this approach is quite arduous: more precisely, it is NP-complete.[26] In complex systems, there are very likely thousands of principles. To guarantee that each equivalence class of principles in every possible combination yields conflict-free sets of permissible options, we would need to solve so-called *satisfiability (SAT) problems*.

SAT is the problem of determining if there exists an interpretation that satisfies a given Boolean formula. It is clear how being a conflict-free set of principles relates to SAT: first, each set of permissible options can be seen as a conjunction of propositions. We add a non-negated proposition a_i if one of the possible options $\phi_i \in \Phi$ is in the set of permissible options, and a negated proposition $\neg a_i$ otherwise. Afterwards, we form a disjunction with each set of permissible options of principles in the same equivalence class. Solving the obtained formula is a classical SAT problem and reveals whether the given sets are conflict-free.

Overall, we need to solve one SAT problem for each equivalence class of principles. SAT is NP-complete, as proven by the Cook–Levin theorem [122]. Consequently, determining whether we can guarantee liveness is practically infeasible. This is even the case when using modern SAT solvers, as there can be many equivalence classes.

Aside from these SAT problems, another problem manifests even earlier, because finding the principles in general is already arduous, even for simple systems. This leads to the question of how this first approach to modeling principles might work for complex systems.

[25] One could also cluster the worlds to specific sets, or use other heuristics, in order to make the evaluation more efficient. We leave all of these considerations for future research.

[26] We will not elaborate on complexity theory here. See [162] for more information.

Medical-Care Robot #25

Asimov's laws serve as an example. Even these seemingly simple and straightforward laws were later supplemented with an additional, zeroth law: "A robot may not harm humanity, or, by inaction, allow humanity to come to harm" [41]. Even today, it is still debated whether introducing such a law makes sense. Essentially, it allows robots to kill humans who infringe on the well-being of humanity.

Should robots be allowed to kill at all? In the principle structure, this law would be the furthest upstream, and consequently supersedes the law stating that a robot is not allowed to harm human beings. However, when exactly is humanity harmed and when is only one person harmed? Are these options different? These questions must be considered when identifying principles, though this is far from easy.

The second approach we suggest fares better with regard to guaranteeing liveness.

7.1.2. Principles as Orders over Options

In our second approach, principles are no longer defined directly. Instead of sets of permissible options, the principles are connected to *option structures* $\langle \Phi, \succeq_\Phi \rangle$, defined in complete analogy to the principle structures introduced in Section 5.3.2, but over the option space Φ. The set of permissible options is, then, determined from these option structures. In the following, we demonstrate a possible means of accomplishing this.

Under the condition that $\omega \models c_\psi$ (for some $\omega \in \Omega$ and some principle $\psi \in \Psi$), we obtain an option structure as described above. This option structure induces a (non-strict weak) *permissibility order* $\succeq_\omega^\psi$, a total preorder on the option set Φ_ω. We simply take the topmost class $[\,\phi\,]$ of this order (in the sense that $\phi \succeq_\omega^\psi \phi'$ for all $\phi' \in \Phi_\omega$) as $Perm^\psi(\omega)$. The intention behind this construction is that the option to perform according to principle ψ in world state ω must be picked from the set highest in the permissibility order associated with that principle.

Naturally, if multiple principles apply for a given state of the world (i.e., more than one c_ψ contains a given world state), the principle highest in the structure $\langle \Psi, \succeq_\Psi \rangle$ is deemed decisive. However, what if different principles in the same (topmost) equivalence class apply? At this point, we deviate from the construction introduced in Section 5.3.3 (which also applies when principles are directly defined). The need for this deviation will become apparent later.

As introduced previously, given a set $\hat{\Psi} \subseteq \Psi$ of principles and an arbitrary world state $\omega \in \Omega$, we refer to the subset of principles that apply to this world state $\{\psi \in \hat{\Psi} \,|\, \omega \models c_\psi\}$ as $\hat{\Psi}^\omega$. Similarly, $O_\omega^{\hat{\Psi}} \coloneqq \{\succeq_\omega^\psi \;|\, \psi \in \hat{\Psi}^\omega\}$ yields the set of relevant option structures, of which each induces a permissibility order $\succeq_\phi \subseteq \succeq_\Phi$ on the option set Φ. We use $\mathbb{O}_\omega^{\hat{\Psi}}$ to denote the set of these orders. Now, if, for a given world $\omega \in \Omega$, it holds that $[\,\psi\,]$ is the topmost class of principles in the principle structure $\mathfrak{P}$, then we are to respect the orders contained in $\mathbb{O}_\omega^{[\psi]}$.

To do so, we use the preorder $\succeq_\omega := \bigcap_{\succeq_\phi \in \mathbb{O}_\omega^{[\psi]}} \succeq_\phi$, obtained by intersecting all element permissibility orders of interest on Φ. Intuitively, we should again take the topmost class of options in the resulting preorder (i.e., the highest equivalence class with respect to $\succeq_\omega$). However, this preorder (as opposed to its constituents) may not have a highest class (owed to the intersection). Thus, while it may not be a weak order itself, it will have maximal classes. One permissive option would be to set $Perm^{\mathfrak{P}}(\omega)$ as the union of all maximal classes of $\succ_\omega$.

Advantages With regard to the properties specified above, this approach demonstrates a decisive advantage when comparing it to the first one: using orders on options makes it easier to guarantee liveness. This is caused by our deviation when trying to identify the set of permissible options with regard to all principles of the same equivalence class. A default option could simply be added relatively far down the option structure of every principle. This option should normally not be executed, but, in a worst-case scenario, it is better than nothing.

By contrast, introducing such an option into the last approach would not lead to an intended outcome, as this would make the fallback option always permissible. Consequently, this option, which was planned only as a fallback, would be executed frequently, which is undesired.

Naturally, we want to avoid introducing fallback options in general. However, if introducing them spares us the (complete) solving of several NP-complete problems, then this could indeed be desirable. Furthermore, this approach is, plausibly, also easy to implement.[27]

In addition to the liveness, which is easier to guarantee, we are not losing anything when using this approach compared to the first one. When assigning the sets of permissible options from the first approach to the highest equivalence class in the option structures here, both approaches coincide. We are strongly inclined to believe that the converse is not possible. Therefore, we could also be gaining, as the other approach could be a subset of this one.

Independently from the consideration of which approach subsumes the other, an important observation can be made. When evaluating the orders over options, one obtains a set of permissible options. Consequently, principles, framed as orders over options, can eventually be translated back into principles as direct implications. This finding does not oppose the finding before it, as we are likely losing some information when the orders are evaluated.

Drawbacks While, from a computational point of view, this approach fares quite well, it has significant drawbacks from a philosophical point of view.

Medical-Care Robot #26

Returning to the example of the robot: at a certain point in time, it faces two option orders from which the set of permissible options must be determined. The first order has “resuscitate the patient” as the most important option, and “comfort the relatives”

[27] We will not go into further detail here, as a thorough description of such an implementation does not deliver any additional value for our purposes.

as the second most important option. The second order has "help the doctors" as the most important option, and "comfort the relatives" as the second most important option. When using the approach described in this section, "comfort the relatives" would be the only permissible option. This result, however, appears counterintuitive at the very least. The robot should generally execute one of the first ranked options in such a situation.

As this example should illustrate, the approach could lead to morally suboptimal results. One might object that these results should be made impossible per construction. Then, however, the advantages of the approach itself are lost, resulting once more in the same problem that principles as direct implications have. Perhaps a $relevance^{\mathfrak{P}}$-function for options could do the trick here. Overall, however, this would once again complicate the implementation.

Even if this problem were solvable, an additional problem remains. The view of principles as orders on options is not an intuitive way of conceptualizing principles. This construction simply neither mirrors any philosophically supported view of principles nor our manner of speaking about them. Consequently, we do not really understand what the principles express when modeling them as orders over options. This is a crucial disadvantage when wanting to express social norms or laws by principles. Doing so easily does not seem to be possible when modeling principles as orders over options.

Setting all problems aside, after presenting the above two approaches to modeling principles, it becomes clear that it is possible to find different formalizations for principles. Each formalization has its own advantages and drawbacks, as discussed above. There are many well-established types of formalism that describe the behavior of systems or agents in some manner. Here, the interesting question arises of whether these formalisms can be productively employed for modeling principles. We believe that this is the case and introduce one such formalism as our next suggestion for modeling principles.

7.1.3. Principles as Deontic Logic Formulae

Our third suggestion is to employ deontic logic formulae. Deontic logic formulae can be used as a means of stating the deontic status of some option(s), and the deontic status of an option describes whether the option is *obligatory*, *permissible*, or *forbidden*. There are a several relations between these statuses: All obligatory options are permissible, but not all permissible options are obligatory. All impermissible options are forbidden. Options that are not explicitly forbidden are permitted, etc. (see Figure 5).

Most systems of deontic logic can be seen as a kind of system of modal logic concerned with the deontic status of options (respectively, actions) and the corresponding relations: is performing the option ϕ allowed/permitted or forbidden/prohibited? Incorporating deontic

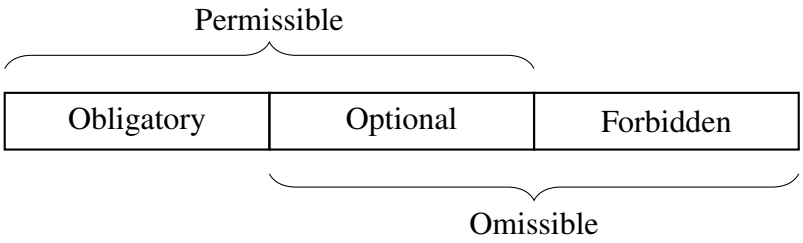


Figure 5: The five deontic statuses of options and how they relate [332].

logic into our approach mirrors the standard manner of conceptualizing normative principles, namely, that of describing permissions and proscriptions. Consequently, this approach already has the advantage of modeling principles naturally and is, therefore, superior to using orders over options, at least in this regard. We want to address the other relevant properties later on.

The language of deontic logic as a whole is very expressive; thus, we must excuse ourselves from a thorough introduction and treatment in this thesis. What we do instead is to offer connection points and ideas, each of them being fruitful for future research. We severely limit the scope of deontic logic in this thesis and caution that, already, such constraint comes with problems. Nevertheless, this should not diminish the value of deontic logic as a whole, as there are many projects that attempt to cope with such problems (see, e.g., [355]). Additionally, deontic logic is, as just indicated, a very natural way of formally expressing normative claims.

In what follows, we will have a look at a very specific sublanguage of standard deontic logic (SDL), the most common version of deontic logic [332]. More specifically, in this approach to modeling principles, a principle links worlds to formulae of a sublanguage $\mathcal{L}$ of SDL, whose evaluation yields the set of permissible options. SDL is one of many deontic logic systems. We do not explore the advantages and disadvantages of different systems; for our purposes, we simply use SDL and leave open the question of whether there are more suitable systems.

Standard Deontic Logic Here are the fundamental building blocks of the deontic logic system that we are employing: The negated performance of an option $\neg perf(\phi)$ is to be read as the omission, the non-performance, of the specific option. Permissibility is expressed by the unary $\mathbf{PE}$ operator. Thus, for options ϕ that are permissible to perform, we write "$\mathbf{PE}perf(\phi)$", and for those that are impermissible to perform, "$\neg\mathbf{PE}perf(\phi)$". The unary operator for obligation ($\mathbf{OB}$) expresses that whatever follows it is obligatory. Since this can be understood to mean that the omission of what follows is impermissible, $\mathbf{OB}$ can be defined as $\neg\mathbf{PE}\neg$. For options ϕ that are obligatory to perform, one can write "$\mathbf{OB}perf(\phi)$".

The language of SDL is classical propositional logic with the operators $\neg$ and $\rightarrow$, supplemented by $\mathbf{OB}$. It has the following axiomatization [332]:[28]

SDL 1 All tautologous well-formed formulae are theorems of SDL.

[28]Note that $\mathbf{PE}$ is not part of the specification of SDL. However, since $\mathbf{PE}perf(\phi) \leftrightarrow \neg\mathbf{OB}\neg perf(\phi)$ is the case, $\mathbf{PE}$ can be introduced without problems.

SDL 2 $\forall \alpha, \alpha' \in \text{SDL}\colon \mathbf{OB}(\alpha \to \alpha') \to (\mathbf{OB}\alpha \to \mathbf{OB}\alpha')$

SDL 3 $\forall \alpha \in \text{SDL}\colon \mathbf{OB}\alpha \to \neg\mathbf{OB}\neg\alpha$.

Additionally, two inference rules are part of SDL: modus ponens, and "if $\vdash \alpha$, then $\vdash \mathbf{OB}\alpha$". The latter has sometimes been termed *the normativity of logical facts*.

We do not present any specific semantic for SDL, as we propose to use a standard Kripke-style semantics [279]. SDL and its standard semantics are much debated, and there are a number of fundamental problems with deontic logics in general, and with SDL in particular. These problems arise primarily because a number of natural sentences must be expressed in SDL in a manner which yields conclusions that seem counterintuitive [332]. However, these are general problems of such formal systems, and they render neither SDL nor any other form of deontic logic useless for expressing propositions about the deontic status of actions or performing options. For now, it suffices that SDL is complete and sound in the standard Kripke-style semantics. As usual, we write $\alpha \models \alpha'$ as a shorthand for "α' is a *semantic consequence* of α", that is, for every interpretation in which α is true, α' is true as well.

For our purposes, we restrict the content of principles to sentences of the following sufficiently rich sublanguage $\mathcal{L}$ of SDL. $\mathcal{L}$ is specified by the following production rules:[29]

$$
\begin{array}{lcrcll}
\alpha & ::= & \beta & \mid & \alpha_1 \wedge \alpha_2 & \\
\beta & ::= & \mathbf{PE}\gamma & \mid & \neg\mathbf{PE}\gamma \quad \mid \quad \beta_1 \vee \beta_2 & \\
\gamma & ::= & \mathit{perf}(\phi) & \mid & \neg\mathit{perf}(\phi) & (\text{for some } \phi \in \Phi)
\end{array}
$$

This language is quite simple and, thus, restrictive. It does not allow, for instance, for conditional obligations and non-normative facts. We acknowledge that more sophisticated models offer additional benefits but leave them for future research.

Nevertheless, several reasons motivated our decision to design $\mathcal{L}$ thusly. First, it is in accordance with the specific manner in which we envision principles, namely, that principles express whether a specific option is or is not allowed. Therefore, our $\mathcal{L}$ includes only the *operator for permissibility* ($\mathbf{PE}$). In addition, the formulae in $\mathcal{L}$ are in conjunctive normal form, which grants them a beneficial structure for evaluation and use.

We allow disjunctions of formulae for two reasons. First, even with this sole addition, this third way of modeling principles subsumes the first one: To express the set of permissible options provided by the first approach by means of deontic logic formulae, it is only necessary to add a conjunction with "$\mathbf{PE}\mathit{perf}(\phi)$" for every ϕ in $\mathit{Perm}(\omega)$. To do so the other way around is not possible because of the disjunctions. Second, allowing disjunctions is already sufficient to illustrate some of the standard problems that arise with respect to deontic logic.

[29] In this thesis, the Backus–Nauer Form is used somewhat liberally. To give an example, non-terminals are identified with derived formulae and indices in the rules. Additionally, brackets will be used, although they are not shown in the grammar. Such simplified notions for grammars are often called *abstract* syntax [50]. We are convinced that this simplification suffices for our purposes.

Medical-Care Robot #27

We now apply this suggestion to the robot. As mentioned previously, $\Phi := \{AnsReq, Charge\}$. A possibility for the robot's principles and their hierarchy is as follows:[a]

1. $prio(req) = H \wedge cost_{task}(req) \leq energy \rightarrow \mathbf{OB}perf(AnsReq)$
2. $prio(req) = L \wedge cost_{task}(req) + dist(CS, req.r) > energy \rightarrow \mathbf{OB}perf(Charge)$

In all other cases, it holds that $\mathbf{PE}perf(AnsReq) \wedge \mathbf{PE}perf(Charge)$, as they are not explicitly forbidden. In this suggestion, it is apparent that even in cases where it is equally likely both principles apply, the robot will try to serve the request.

[a] We used the obligation operator here (which is not in $\mathcal{L}$). Doing so poses no problem, as $\mathbf{OB}\phi$ is equivalent to $\neg\mathbf{PE}\neg\phi$ (as already stated) and $\neg\mathbf{PE}\neg\phi$ is part of $\mathcal{L}$. In addition, we use implications in order to express the association of formulae to worlds (which is somewhat unusual).

Determining the Set of Permissible Options To determine the set of permissible options from these formulae, we begin with the set of permissible options given by a single principle and a world in which this principle applies. As previously mentioned, we associate a formula α to each principle ψ (at least in this third approach to modeling principles), expressing the permissions and proscriptions of this principle.

To determine the set of permissible options in a world $\omega \in \Omega$ given by the principle $\psi \in \Psi$ (which applies in ω), we check for every option $\phi \in \Phi_\omega$ whether the permissibility of performing this option is compatible with the formula $\alpha \in \mathcal{L}$ implied by ψ together with $\overline{\omega}$. Given some principle $\psi \in \Psi$ and a world state $\omega \in \Omega$, determining the set of permissible options resembles the following:

$$\forall \omega \in \Omega, \forall \psi = (\overline{\omega} \rightarrow \alpha) \in \Psi^\omega, \forall \phi \in \Phi_\omega: \quad (\overline{\omega} \wedge \psi) \rightarrow (\phi \in Perm^\psi(\omega) \leftrightarrow (\alpha \wedge \mathbf{PE}perf(\phi) \not\models \bot))$$

By using the same approach as in Section 5.3.3, the set of permissible options with regard to all principles can be derived, which is, naturally, only important in cases of perfect knowledge. To determine the set of permissible options, it is necessary to solve a SAT problem (just as with our first proposal). The formulae in $\mathcal{L}$ are in disjunctive normal form, and we have to check whether they are satisfiable to arrive at the set of permissible options. Therefore, we have at least one problem with regards to runtime when implementing the approach.

At this point, we first note that an additional axiom is needed when using SDL, which is owed to our subscription to liveness (Axiom 2). This need arises because there is a practical problem with SDL in some cases, other than the theoretical problems already mentioned.

Medical-Care Robot #28

In a situation where the robot is in a room with a dying patient, the robot cannot be obliged to resuscitate the patient if it is not able to do so, for instance, if it does not have sufficient energy. If the robot were nevertheless obliged to do so, this would undermine the concept of obligation. In this case, the robot is, thus, not only permitted not to resuscitate the patient, but it could even be permitted to return to the charging station.

This restriction is often called "ought implies can" and claims that, if an agent is morally obliged to perform a certain option, then it must logically and maybe even (meta)physically be able to perform it. The first formulation of this restriction is often ascribed to Immanuel Kant. He notes that: "The action to which the 'ought' applies must indeed be possible under natural conditions." [260, A548/B576]. "Ought implies can" has two implications for our framework. The first, direct implication is that no principle should demand an option that is not available. This can be incorporated into a new axiom:

Axiom 3 $\forall\omega \in \Omega, \forall\psi = (\overline{\omega} \to \alpha) \in \Psi^{\omega}, \forall\phi \in \Phi\colon \quad (\overline{\omega} \wedge \psi) \to ((\alpha \models \mathbf{OB}\mathit{perf}(\phi)) \to \phi \in \Phi_{\omega})$ **[Ought → Can]**

The second implication is that, because of Axiom 1, we know that options are distinct. Together with Axiom 3, we can deduce the following statement in our framework:

Lemma 2 $\forall\phi, \phi' \in \Phi\colon \mathbf{OB}\mathit{perf}(\phi) \wedge \mathbf{OB}\mathit{perf}(\phi') \to \phi = \phi'.$

To put this statement into words: at all moments in time, there can only be precisely one option that a system is obliged to carry out. This is only natural, as a system cannot be obliged to simultaneously perform two different actions, for this is logically and maybe even (meta)physically impossible, assuming distinct options.[30]

After clarifying the need for this axiom, we turn back to the disjunctions. As stated above, allowing disjunctions to be in $\mathcal{L}$ serves as an example to illustrate some problems which arise even with this simple sublanguage of SDL. This is partly so because several questions remain open, including the issues of what it means for several options to be disjunctively permitted, whether their permissibility is divided between them or not. These questions are difficult to answer and controversially discussed [332]. Additionally, it is still not clear how to make a final choice regarding the options that are permissible to perform.

A much-debated problem arising due to disjunctions in deontic logic is *Ross' Paradox* [401]. In the light of this problem and the questions above, it appears hard to find a satisfying way to interpret disjunctive deontic logic formulae. The even more expressive SDL, then, has

[30] This is only true for *actions*. If we are concerned with *states of affairs*, it could very well be possible for more than one state to be obliged to be the case at a time. At all times, there is an obligation that no one is tortured, and there is also an obligation that animals are kept in a manner appropriate to their species, etc.

even more problems. Thus, it appears that we are forced to abandon the approach of modeling principles as as deontic logic formulae.

This seems to be the case at least when directly implementing the framework as we have outlined it above. However, in some cases it might be beneficial to deviate in some aspects if this allows us, for instance, to model the principles in a more appropriate way. In a previous work, we have shown how this might look for deontic logic formulae [446].

7.2. An Alternative Formalization of the Framework

In this thesis, we will demonstrate a different approach. Generally, it stands to reason that a formal framework for machine ethics is based on a logic for actions. After all, in order to formally specify the normative constraints on a machine's behavior, one needs a system that provides the right kind of formal surface: as norms involve specific properties of agents, actions and outcomes, we need a logical system that involves these components.

With our framework, we have described the most basic components that such a system would require. As a final part of substantiating our framework, we will show that established logical systems can be leveraged to flesh it out (see [446] for our previous approach for this). As this is not a key contribution of our thesis, however, this fleshing out will remain approximate.

7.2.1. STIT, XSTIT and Beyond

As the family of so-called STIT logics ("STIT" is the acronym for "**S**ees **T**o **I**t **T**hat") allows direct discussion of choice-making, these logics are especially well-suited for machine ethics. First, because they have been thoroughly studied. Second, because they have been proven to be highly flexible. For instance, STIT semantics have been proposed as a version of two-dimensional semantics forming an alternative to traditional Kripke-style possible world semantics for a number of systems of modal logic such as deontic logic, epistemic logic and temporal logic. These are all modal systems that will be needed in a full-blown machine-ethics framework [242, 346]. Accordingly, we build our alternative approach to such a framework on top of STIT-related work of John Horty [238, 240] and Jan Broersen [85, 86].

STIT semantics are grounded in a philosophical theory of indeterministic time which was first set out by Arthur Prior [380], and refined by Richmond Thomason [461]. The result of considering the full temporal evolution of a world is a series of ordered moments, from the past to the present to the future, called a *history*. Indeterminism comes into play insofar as things could have come about differently at any moment. Therefore, each moment is part of several distinct histories, all of which coincide up to that single moment and branch off at that point, or a later moment.

In this thesis, we will restrict ourselves to the case of only an individual agent acting (as opposed to sets of agents), and, for this reason, fall short of the semantics proposed by, for

instance, Broersen [85]. Beyond that, we adopt Broersen's notation as opposed to Horty's.[31] Furthermore, we choose a STIT logic where the effects of actions take place in the very ne**X**t moment (therefore called **X**STIT logic) and not in the same moment the action is performed. This has a couple of theoretical advantages. For instance, it makes a multi-agent extension axiomatizable and decidable.

As we are concerned with the morally constrained interplay between a machine and its environment, we are not using XSTIT, but an extended version of it which allows for such an interplay. First, we add subjective probabilities from Broersen [86].[32] To this end, we add a second agent (the world) with which a machine must interact and about whose state and behavior, for the most part, only probabilities are known. Moreover, we add the obligation and permissibility operators as suggested by Broersen [85]. Finally, we add action types, an idea proposed by Horty [240]. Overall, this results in the language AT-XSTIT^p (**A**ction **T**ype XSTIT with **p**robabilities), which we will outline below.[33]

We first introduce the semantics for a probabilistic version of XSTIT and add action types and obligation later on. As a general preparation, we introduce the following notations:

- M is an infinite, but countable, set of *moments*. Elements of M are called moments and are denoted by m, m', ...
- $<_M \subseteq M \times M$ is a serial and strict partial order with no backward branching, that is: $\forall m, m', m'' : (m' <_M m \wedge m'' <_M m) \rightarrow (m' = m'' \vee m' <_M m'' \vee m'' <_M m')$.
- We define the set of histories $H \subseteq M^\omega$ as the largest set of words over M respecting the order $<_M$. We use h, h', h_0,... to denote histories, and we use a couple of history-related abbreviations:
  - We write $m \in h$ in order to express that there is an occurrence of m in h, that is, there are (left/right) infinite words w and w' so that $h = wmw'$. Note that, by definition of $<_M$, each $m \in h$ occurs at most once in h.
  - Whenever we have a $m \in h$ as reference, we will use $\mathbb{Z}$ as an index set centered at m. In other words, given a moment m and a history h with $m \in h$, we write $h[0] = m$. Given a moment m and an $i \in \mathbb{Z}$, we write $h[i \ldots]$ in order to denote the infinite suffix of h starting at $h[i]$, that is, $h[i \ldots] = h[i]h[i+1]h[i+2] \ldots$.
  - For every moment $m \in M$, we define $H^m \subseteq H$ as the set of all histories h containing the corresponding moment m. Formally expressed: $H^m := \{h \in$

[31] We do so because Broersen's notation is closer to the notations used in traditional computer science. In contrast, Horty's notation conforms more to the way philosophers think about STIT semantics. While this is a mainly philosophical thesis, the people who will need to work with the proposed formalizations are presumably computer scientists. For this reason, we have chosen the formalization that seems more inclined towards application.

[32] We leave the question of incorporating non-determinism into the model for future work.

[33] To increase readability, we assume universal quantification over unbound variables in what follows.

$H \mid m \in h\}$. Note that by the nature of $<_M$, the histories in H^m share the same left infinite prefix ending in m.

- We use I to specify elements in H^m by defining $I \subseteq M \times H$ such that $I := \{\langle m, h\rangle \in M \times H \mid m \in h\}$. We call I the set of indices, and write $M_{\mathcal{M}}$ and $I_{\mathcal{M}}$ whenever we see the need to emphasize that we refer to the set of moments or indices, respectively.
- We define $succ : I \mapsto M$ as the function that yields the "next moment" for a given index $\langle m, h\rangle \in I$, i.e., the unique next moment m' following m in h, i.e., $h = wmm'w'$ for some (left/right) infinite words w and w'.

XSTIT^p With all of the above, we can now define the language of XSTIT with subjective probabilities (i.e., the language of XSTIT^p):

Definition 7 Given a countable set of propositions $Props$ and $p \in Props$, and a set of two agents $Agents = \{Sys, Wld\}$ and $\alpha \in Agents$, the formal language $\mathcal{L}_{\text{XSTIT}^p}$ is:

$$\varphi \coloneqq p \mid \neg\varphi \mid \varphi \wedge \varphi' \mid \Box\varphi \mid [\alpha \text{ xstit}]\,\varphi \mid X\varphi$$

Besides the usual connectives known from propositional logic, the syntax of XSTIT includes three modal operators. The $\Box\varphi$ operator expresses "historical necessity". In other words, it expresses that all possible histories containing a certain moment m have the property φ at m. In general, this is interpreted as a universal quantification over the branching dimension of time. The operator $[\alpha \text{ xstit}]\,\varphi$ stands for "agents α sees to it that φ in the next state". The last modal operator is the *next* operator $X\varphi$. It expresses the transition to a next static state. [85]

Definition 8 A XSTIT^p-frame is a tuple $\langle M, <_M, E, B_{Sys}\rangle$ such that:

- M and $<_M \subseteq M \times M$ are defined as above.
- $E : I \times Agents \mapsto 2^M \setminus \emptyset$ is an effectivity function yielding for an agent $\alpha \in Agents$ the set of next moments that may follow the agent's actions relative to an index $\langle m, h\rangle \in I$. E has to fulfill the following constraints:
  - $succ(\langle m, h\rangle) \in E(\langle m, h\rangle, \alpha)$;
  - if $m' \in E(\langle m, h\rangle, \alpha)$ then $\exists h' \in H^m : m' = succ(\langle m, h'\rangle)$;
  - if $m' = succ(\langle m, h\rangle)$ and $\exists h' \in H^m : h' \in H^{m'}$ then $m' \in E(\langle m, h'\rangle, \alpha)$;
  - $\forall h, h' \in H^m : E(\langle m, h\rangle, Wld) \cap E(\langle m, h'\rangle, Sys) \neq \emptyset$.

  For every moment $m \in M$ and agent $\alpha \in Agents$, E induces a partition $Choices^{m,\alpha}$ of H^m. We call the elements $Ch \in Choices^{m,\alpha}$ *choices* and for $h \in H^m$ we write $Choices^{m,\alpha}(h)$ for the partition block containing the history h. The history $h \in H^m$

of a moment $m \in M$ is in block $Choices^{m,\alpha}(h')$ if and only if $E(\langle m, h\rangle, \alpha) = E(\langle m, h'\rangle, \alpha)$. We let $Choices^{\mathcal{M},\alpha}$ denote $\bigcup_{m\in M} Choices^{m,\alpha}$ for some agent α.

- $B_{Sys} : M \times Choices^{\mathcal{M},Wld} \mapsto [0, 1]$ is a subjective probability function for agent Sys such that $B(m, Ch)$ (for $Ch \in Choices^{m,Wld}$ being one of the choices Wld can choose in m) expresses agent Sys's belief that in moment m, agent Wld performs that choice. We apply the following constraints:
  - $B_{Sys}(m, Ch) = 0$ if $Ch \notin Choices^{m,Wld}$;
  - $B_{Sys}(m, Ch) \geq 0$ if $Ch \in Choices^{m,Wld}$;
  - $\sum_{Ch\in Choices^{m,Wld}} B_{Sys}(m, Ch) = 1$.

We call the elements of $Choices^{m,\alpha}$ the *action tokens* available at index $\langle m, h\rangle$ to agent $\alpha \in Agents$. We say that an action token Ch is *associable* to a history h if and only if $h \in Ch$.

If we want to evaluate propositions in our frame, we need a model:

Definition 9 A XSTIT^p-frame $\mathcal{F} = \langle M, <, E, B_{Sys}\rangle$ is extended to an XSTIT^p-model $\mathcal{M} = \langle M, <, E, B_{Sys}, \pi\rangle$ by adding a valuation π of atomic propositions:

- π is a valuation function $\pi : Props \to 2^I$ assigning, to each atomic proposition, the set of indices in which it is true.

The truth conditions for the semantics of the operators are standard [85]. The non-standard aspect is the two-dimensionality of the semantics [85], meaning that we evaluate truth with respect to indices built from a dimension of histories and a dimension of static states.

Definition 10 Given a XSTIT^p-model $\mathcal{M} = \langle M, <_M, E, B_{Sys}, \pi\rangle$, truth at an index $\langle m, h\rangle$ is defined as:

$\langle m, h\rangle \models p \Leftrightarrow \langle m, h\rangle \in \pi(p)$
$\langle m, h\rangle \models \neg\varphi \Leftrightarrow$ not $\langle m, h\rangle \models \varphi$
$\langle m, h\rangle \models \varphi \wedge \varphi' \Leftrightarrow \langle m, h\rangle \models \varphi$ and $\langle m, h\rangle \models \varphi'$
$\langle m, h\rangle \models \Box\varphi \Leftrightarrow h' \in H^m$ implies $\langle m, h'\rangle \models \varphi$
$\langle m, h\rangle \models X\varphi \Leftrightarrow succ(\langle m, h\rangle) = m'$ implies $\langle m', h\rangle \models \varphi$
$\langle m, h\rangle \models [\alpha \text{ xstit}]\,\varphi \Leftrightarrow h' \in Choices^{m,\alpha},\ m' = succ(\langle m, h'\rangle)$ implies $\langle m', h'\rangle \models \varphi$

Obligation and Permissibility We now extend our framework by two central deontic operators: a modal operator $\mathbf{OB}$ for "ought to do" and a modal operator $\mathbf{PE}$ for "permissible to do". We have already discussed the relation between central deontic concepts above (in Section 7.1.3) and visualized it in Figure 5.

The operators $\mathbf{OB}$ and $\mathbf{PE}$ are implemented via the introduction of so-called *violation constants*, as proposed by Bartha for traditional STIT logic [56], and by Broersen for XSTIT [85]. The underlying idea for this construction was first proposed by Anderson [19]. Violation constants represent actions that ought not to be performed.

We follow Broersen in writing v for the violation constant that is added to the set of propositional variables: $v \in Props$ [85]. This will be the connection point for our moral principles; it is the place where, later, our moral principles interact with the models of $\mathcal{L}_{\text{AT-XSTIT}^p}$.

Definition 11 The operator $\mathbf{OB}\,[\alpha \text{ xstit}]\,\varphi$, expressing the *obligation* of α to see to it that φ, is defined as follows:

$$\mathbf{OB}\,[\alpha \text{ xstit}]\,\varphi \equiv_{def} \Box\,(\neg\,[\alpha \text{ xstit}]\,\varphi \rightarrow [\alpha \text{ xstit}]\,v)$$

In natural-language terms, the above construction states that it is obligatory for an agent to do something if and only if by not doing it, they perform a violation. Since the effect of the obliged action can only be felt in subsequent states, violations have to be properties of subsequent states.

Proposition 1 The operator $\mathbf{OB}\,[\alpha \text{ xstit}]\,\varphi$ is KD[34], that is, it has the same properties as standard deontic logic. (A proof of this proposition can be found in [85].)

Definition 12 The operator $\mathbf{PE}\,[\alpha \text{ xstit}]\,\varphi$, expressing *permissibility* for α to see to it that φ, is defined as follows:

$$\mathbf{PE}\,[\alpha \text{ xstit}]\,\varphi \equiv_{def} \Box\,([\alpha \text{ xstit}]\,\varphi \rightarrow \neg\,[\alpha \text{ xstit}]\,v)$$

Once again, let us express the above construction in natural terms. It is permissible for an agent to do something if and only if by doing it, they do not perform a violation.

Many frameworks of STIT logics stop at this point. We regard that as unsatisfactory, as it is not clear, at this point, whence the violation constants v are derived. In order to incorporate them satisfactorily, we need *action types* and *principles*. Our first focus is on action types.

AT-XSTIT^p In order to introduce action types into XSTIT^p, we define a countable set of action-type labels $ActionTypes = \{\tau_1, \tau_2, \ldots\}$ and a function for a given XSTIT^p-model $\mathcal{M}$, assigning labels to action tokens (given a particular moment).

Definition 13 Given a XSTIT^p-model $\mathcal{M}$, we let $Label^{\mathcal{M}} : Choices^{\mathcal{M}} \mapsto ActionTypes$ be a labeling function, mapping action tokens to action types.

[34]The logic K emerges by adding an operator $\mathbf{OB}$ and two principles to propositional logic: 1) If A is a theorem of K then so is $\mathbf{OB}A$ (rule of obligation), and 2) $\mathbf{OB}(A \rightarrow B) \rightarrow (\mathbf{OB}A \rightarrow \mathbf{OB}B)$ (distribution axiom). We arrive at KD by adding the D axiom: $\mathbf{OB}A \rightarrow \mathbf{PE}A$. [186]

We call XSTIT^p with action types AT-XSTIT^p. An AT-XSTIT^p frame is an XSTIT^p-frame extended by the labeling function: $\mathcal{F} = \langle M, <_M, E, B_{Sys}, Label\rangle$. The model is defined similarly to an XSTIT^p model. We use $\mathfrak{M}$ to denote the space of models of $\mathcal{L}_{\text{AT-XSTIT}^p}$.

7.2.2. Moral Principles

We now introduce formal moral principles into our framework. They determine the moral constraints to which our system is subjected, by placing the violation constants into a model. For this purpose, principles are considered to be CTL^* formulae over the set $ActionTypes$ and the language $\mathcal{L}_{\text{AT-XSTIT}^p}$.

CTL^* is a very expressive temporal logic which is decidable for finite-moment structures and used frequently in verification [169]. Consequently, using CTL^* for our principles is a first step towards the verifiability of our framework.[35] How to spell out this verifiability in detail, however, must be left to future work.

Let $\Psi = \{\psi_1, \ldots, \psi_m\}$ be the set of principles once more. The general idea is that principles determine the violations of norms; thus, they place violation constants into models. For this, one checks whether a history satisfies the principle-induced CTL^* property, and places a violation constant wherever the property is violated.

We begin by defining the syntax of our variant of CTL^*, which we call $CTL^*_{\text{AT-XSTIT}^p}$

Definition 14 We define $CTL^*_{\text{AT-XSTIT}^p}$ over $ActionTypes$ and $\mathcal{L}_{\text{XSTIT}^p}$. In what follows, we use τ to range over sets of $ActionTypes$ and φ for atomic propositions from $\mathcal{L}_{\text{XSTIT}^p}$.
$CTL^*_{\text{AT-XSTIT}^p}$ *formulae* are defined as:

$$\rho ::= \varphi \,|\, \neg\rho \,|\, \rho_1 \wedge \rho_2 \,|\, \exists\rho \,|\, \bigcirc_\tau \rho \,|\, \rho_1 \mathbf{U} \rho_2$$

Definition 15 Given an AT-XSTIT^p-model $\mathcal{M} = \langle M, <_M, E, B_{Sys}, Label, \pi\rangle$, an index $\langle m, h\rangle$ satisfies a $CTL^*_{\text{AT-XSTIT}^p}$ formula according to the minimal relation $\models^*$ satisfying the following constraints:

$\langle m, h\rangle \models^* \varphi \Leftrightarrow \langle m, h\rangle \models \varphi$
$\langle m, h\rangle \models^* \neg\rho \Leftrightarrow$ not $\langle m, h\rangle \models^* \rho$
$\langle m, h\rangle \models^* \rho_1 \wedge \rho_2 \Leftrightarrow \langle m, h\rangle \models^* \rho_1$ and $\langle m, h\rangle \models^* \rho_2$
$\langle m, h\rangle \models^* \exists\rho \Leftrightarrow$ there exists $Ch \in Choices^{m,Sys}$ with $h' \in Ch$ and $\langle m, h'\rangle \models^* \rho$
$\langle m, h\rangle \models^* \bigcirc_\tau \rho \Leftrightarrow \langle succ(m, h), h\rangle \models^* \rho$ and $Label(Choices^{m,Sys}(h)) \in \tau$
$\langle m, h\rangle \models^* \rho_1 \mathbf{U} \rho_2 \Leftrightarrow \exists n \geq 0 : \langle succ^n(m, h), h\rangle \models^* \rho_2$
and $\forall 0 \leqslant k < n : \langle succ^k(m, h), h\rangle \models^* \rho_1$

In the same way (and for the same reason that principles can sometimes be outweighed) as in Section 5.3.2, we define $\succeq_\Psi \subset \Psi \times \Psi$: a total preorder on the set of principles.

[35] A future approach could attempt to express the whole semantics as CTL^*, or adopt Broersen's approach of a combination of CTL and STIT [84]. We refrain from doing so (yet), as STIT semantics are currently far better grounded for expressing the actions a system can and should make.

In order to adjust the violations of this order over principles, we not only assume one violation constant $v \in Props$, but also a countably infinite number of constants v_i^k for each equivalence class Ψ_i of principles in a set of principles Ψ and for all $k \in \mathbb{N}$. We construct an order $\succeq_{\mathcal{V}}$ over these constants mirroring $\succeq_{\Psi}$, i.e., for all $i, j \in \{1, \ldots, t\}$. It holds true that $\forall k, l \in \mathbb{N} : v_i^k \succeq_{\mathcal{V}} v_j^l$ if and only if $\Psi_i \succeq_{\Psi} \Psi_j$. We call the set of these variables $\mathcal{V} \subseteq Props$. In order to do justice to this change, we redefine the $\mathbf{OB}$ operator:

Definition 16 We define n ought-to-do operators $\mathbf{OB}_1, \ldots, \mathbf{OB}_n$ with:

$$\mathbf{OB}_i\,([\alpha \text{ xstit}]\varphi) \text{ if and only if } \exists k > 0 \in \mathbb{N} : \Box\left(\neg\,[\alpha \text{ xstit}]\,\varphi \rightarrow [\alpha \text{ xstit}]\,v_i^k\right)$$

We derive, analogously to Definition 12, n permissible-to-do operators $\mathbf{PE}_1, \ldots, \mathbf{PE}_n$.

With the construction thus far, we can define the following function, $Apply_{\mathcal{M}}$ for applying a principle in a certain model $\mathcal{M}$.

Definition 17 Given some model AT-XSTIT^p-model $\mathcal{M} = \langle M, <_M, E, B_{Sys}, Label, \pi\rangle \in \mathfrak{M}$, the function $Apply_{\mathcal{M}} : \Psi \times I_{\mathcal{M}} \mapsto \mathfrak{M}$ yields, for a principle ψ and an index i, a possibly modified version $\mathcal{M}'$ of $\mathcal{M}$ that differs from $\mathcal{M}$ maximally insofar as it adds a violation constant to indices of histories which violate ψ. More exactly, for $\psi \in \Psi_i$:

$$Apply_{\mathcal{M}}(\psi, \langle m, h\rangle) = \begin{cases} \langle M, <_M, E, B_{Sys}, Label, \pi'\rangle & \text{if } h \not\models^* \psi \\ \mathcal{M} & \text{otherwise} \end{cases}$$

with π' differing from π only insofar as: $\forall \varphi \in \mathcal{L}_{\text{AT-XSTIT}^p}$: for all $h' \in H^m$: if $\mathcal{M}, \langle m, h'\rangle \models [Sys \;\; \text{xstit}]\varphi$, then (for $m' = succ(\langle m, h'\rangle)$): if $\langle m', h'\rangle \in \pi(v_i^k)$ for some $k \in \mathbb{N}$ we set $\langle m', h'\rangle \notin \pi'(v_i^k)$ and $\langle m', h'\rangle \in \pi'(v_i^{k+1})$.

In other words, $Apply_{\mathcal{M}}$ increases the violation counter corresponding to a principle if that principles' property is not satisfied by a history. In order to understand $Apply_{\mathcal{M}}$ better, we introduce a proposition: If a principle ψ is not satisfied at index $\langle m, h\rangle$ in model $\mathcal{M}$, then everything Sys is permitted to see to it to happen at this index is impermissible for the agent to see to it that it happens in the model $\mathcal{M}'$ resulting from the application of the principle:

Proposition 2 For $\mathcal{M}' = Apply_{\mathcal{M}}(\psi, \langle m, h\rangle)$ (for a principle $\psi \in \Psi_i$ and index $\langle m, h\rangle$): if $\mathcal{M} \neq \mathcal{M}'$ then for all $\varphi \in \mathcal{L}_{\text{AT-XSTIT}^p}$ for which $\mathcal{M}, \langle m, h\rangle \models [Sys \;\; \text{xstit}]\varphi$:

$$\mathcal{M}', \langle m, h\rangle \models \neg\mathbf{PE}_i([Sys \;\; \text{xstit}]\varphi).$$

Note that it trivially holds that, if seeing to it that φ is permissible for Sys at some index $\langle m, h\rangle$ on some model $\mathcal{M}$ and ψ is satisfied by h in this model, then seeing to it that φ remains

permissible for Sys even in the model resulting from the application. After all, if the principle ψ is satisfied by h, the result of $Apply_{\mathcal{M}}$ is $\mathcal{M}$.

Based upon $Apply_{\mathcal{M}} : \Psi \times I \mapsto \mathfrak{M}$, we introduce $ReApply_{\mathcal{M}} : 2^{\Psi} \times I \mapsto \mathfrak{M}$ as follows:

Definition 18 Given some AT-XSTIT^p-model $\mathcal{M} = \langle M, <_M, E, B_{Sys}, Label, \pi \rangle \in \mathfrak{M}$, the function $ReApply_{\mathcal{M}} : 2^{\Psi} \times I_{\mathcal{M}} \mapsto \mathfrak{M}$ for a set of principles $\Psi = \{\psi_1, \ldots, \psi_r\} \in 2^{\Psi}$ and an index i yields a possibly modified version $\mathcal{M}'$ of $\mathcal{M}$ by applying $Apply$ recursively:

$$ReApply_{\mathcal{M}}(\Psi, i) = Apply_{\mathcal{M}_{r-1}}(\psi_r, i), \text{ where } \mathcal{M}_{r-1} = Apply_{\mathcal{M}_{r-2}}(\psi_{r-1}, i) \text{ and } \mathcal{M}_0 = \mathcal{M}$$

The function $ReApply$ steps recursively through all principles in a set Ψ and applies them one after another to an initial model $\mathcal{M}$. Finally, because a system navigating through the moments of a frame through time does not know the history in which it is, we need a function $BlindApply_{\mathcal{M}}$ that only operates on a moment m and not on a complete index:

Definition 19 Given some AT-XSTIT^p-model $\mathcal{M} = \langle M, <_M, E, B_{Sys}, Label, \pi \rangle \in \mathfrak{M}$, the function $BlindApply_{\mathcal{M}} : 2^{\Psi} \times M_{\cdot} \to \mathfrak{M}$ for a set of principles $\Psi = \{\psi_1, \ldots, \psi_r\} \in 2^{\Psi}$ and a moment m yields a possibly modified version $\mathcal{M}'$ of $\mathcal{M}$ by applying $ReApply_{\circ}$ recursively to every $h \in H^m$.

Obviously, the fact that H^m is countably infinite by construction makes a realization of the function practically and computationally infeasible, but this should not bother us in the context of our theoretical framework.

We call $\mathfrak{P} := \langle \Psi, \succeq_{\Psi}, BlindApply_{\cdot} \rangle$ a *principle structure*, giving us a hierarchy of moral principles flexible enough to also accommodate the absence of any hierarchy (obtained if $\approx_{\Psi} := \Psi \times \Psi$) and application function which requires parameterization with a model.

At this point, we have filled our alternative framework to the point where it is equivalent to our original framework up to the introduction of explanations (i.e., everything we outlined in Section 5). To this end, we have used proven logical systems to arrive at the same result.

All in all, this should be a first indicator that our framework is not a mere figment of our imaginations. We will evaluate the second part of the framework, namely, the explanations it potentially generates, later (in Section 11). At this point, however, we are finished with elaboration on our framework, and come to the topic of machine explainability in more detail.

Part III.
Machine Explainability

8. Charting the Field of (Machine) Explainability

At this point, we have motivated the need for machine ethics, and have shown how it could be possible to implement it into a system. Furthermore, we have also motivated the need for machine explainability based on machine ethics. However, we remained silent on what exactly machine explainability is, and on how our approach specifically supports machine explainability. This part of the thesis is dedicated to amending the first of these two deficits.

8.1. What are Explanations?

When talking about machine explainability, one must first be clear about what explanations are, as such. Explanations themselves have a long history in philosophy. Aristotle, for instance, provided an account of what he believed explanations are, namely, answers to "why"-questions [34]. Even today, the more general view that explanations are answers to "w"-questions (e.g., "why", "what", "when", but also "how"[36]) is widely spread (see, e.g., [131, 188, 206]) [107].

Apart from this initial account, Aristotle also proposed a first distinction: One must differentiate those answers to "why"-questions that provide only *symptoms* of something being the case, from those that provide *reasons* for it being the case [356].[37] If one wants to know why somebody is ill, one can either answer with something like "There are symptoms such as fever, increased sweating, . . . " or with something like "The person in question returned from Asia a few days ago, where he was bitten by a mosquito, so it is likely that he is infected by malaria". Later, approximately during the lifetime of Kant, the terms *rationes cognoscendi* for the first variety of answers, and *rationes essendi* for the second variety, were coined.

More than 2000 years later, there are still many competing conceptions of what explanations are. Just as with the debate over the "correct" normative theory, there is no consensus on a "correct" account of explanation. What is considered an explanation can sometimes vary to such a great extent that there is no single concept of explanation, but rather a variety of concepts linked by family resemblance. Thus, exactly defining what characterizes an explanation is difficult. Nevertheless, this will not deter us from charting the field.

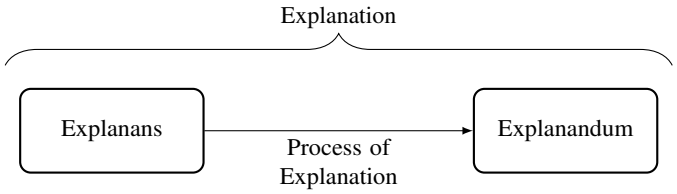


Figure 6: The general model of explanation [171].

[36] By using the archaic interrogative word "wherefore", one can even have them all start with a "w".

[37] To be exact, Aristotle differentiated *three* varieties of answers. The third variety (later called *rationes fiendi*), however, is of no interest to us. Those answers are concerned with the outer influences which led to something being the case. We, in contrast, are mostly concerned with the internal reasoning of a system.

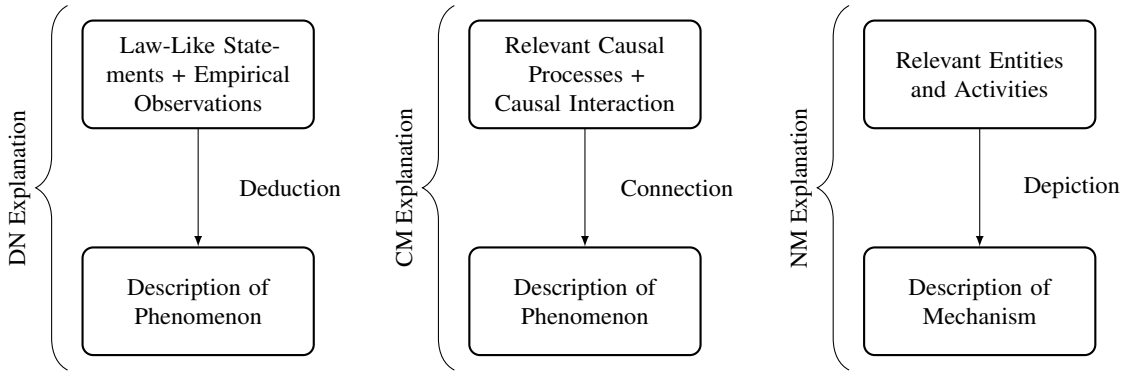
In general, it is uncontroversial that explanations consist of two main components: the *explanandum* and the *explanans*. The explanandum is the phenomenon to be explained. The explanans, on the other hand, takes on the explanatory role and is usually the set of propositions cited within an explanation to explain the explanandum [363]. This general scheme is visualized in Figure 6. In what follows, we will call the inference from explanans to explanandum the *process of explanation*, and the product of this inference the *explanation*.

8.1.1. Scientific Explanation

Since explanation is, among others, deeply rooted in science, a popular way to approach explanation is through science. Disciplines as diverse as psychology, sociology, and biology have one thing in common: they aim to explain certain phenomena (e.g., why an event occurred or why a fact exists). The variety of explanation accounts is partly due to disciplinary diversity. Because different disciplines have their own distinct ways of explaining things, there are many different accounts of explanation [364]. In what follows, we will briefly introduce some of these accounts.

The branch of philosophy that deals with explanation in science (and, more generally, with the conditions under which a research discipline qualifies as a science at all) is the *philosophy of science*. It is not far-fetched to say that, for many scholars, one of the goals of philosophy of science is to to find a single account of explanation that adequately captures the different modes of explanation in the various sciences [500].

In this respect, we can observe an interesting commonality with practical philosophy: while one goal of practical philosophy is to find a universal normative theory, one goal of philosophy of science is to find a universal account of explanation. However, a look at the literature shows that, among the various accounts of scientific explanation, no single one is currently preferred. Rather, there are several accounts that are seen as complementary [324, 364, 419, 500].



(a) The model of deductive-nomological explanation (DNE).

(b) The model of causal-mechanical explanation (CME).

(c) The model of new mechanist explanation (NME).

Figure 7: Schematic overview of different accounts of explanations [171].

Among the various accounts of scientific explanation, we will discuss three in particular: *DNEs*, *CMEs*, and *NMEs*. The general scheme of these three accounts is visualized in Figure 7. The DNE is the oldest type of dedicatedly scientific explanation, and was proposed by Hempel and Oppenheim in the first half of the twentieth century [217, 218]. A DNE is primarily a deductive argument, where the conclusion is the explanandum and the premises contain at least one regularity or law of nature (see Figure 7a) [363]. Although Hempel also intended DNEs to be used to deduce special laws from more general ones, they are generally employed to explain specific events and to make predictions about the state of a system [356]. Because of their focus on laws, DNEs are often referred to as *covering law explanations*.

DNEs can be considered the first "real" account of scientific explanation, and subsequent accounts have often been devised to improve on the problems of this account. For example, Salmon developed the statistical relevance account to address failures of DNEs to capture information about relevant causes [406]. He then discarded this approach and devised CMEs, still with the goal of capturing the allegedly central role of causation in explanation [407, 408]. A similar motivation stands behind the related account of Dowe [158].

Overall, Salmon was concerned with an analysis of causality that is compatible with our knowledge of physics. To this end, he distinguished two aspects of causation: *causal processes* and *causal interactions*. A causal process is the ability to transfer a mark or its own physical structure in a spatio-temporally continuous manner (e.g., the motion of sound waves through air). A causal interaction takes place when causal processes interact and modify their structure (e.g., the interference of sound waves). A successful CME involves citing some parts of the causal processes and interactions that led to the phenomenon in question (see Figure 7b).

However, CMEs have also not escaped criticism. Most prominently, both DNEs and CMEs are criticized for not adequately capturing the explanatory practices or needs of the special sciences. Because DNEs and CMEs are tailored to the explanatory needs of physics, some philosophers found that these accounts of explanation do not meet the needs of other sciences. For instance, it is by no means clear that laws of nature play the same central role in biology as they do in physics [65, 364, 437]. Moreover, Hitchcock noted problems with the kind of causality that is captured by CMEs [226]. While CMEs may be successful in explaining simple sound wave interference, it is difficult to see how they apply to the tangled causal networks studied in the special sciences. As a result, new explanation accounts have been developed, such as NMEs, causal interventionist explanations, and network explanations.

The idea behind NMEs is that providing an explanation involves demonstrating how some phenomena arise from a collection of entities and activities (see Figure 7c). A successful explanation involves identifying the entities and activities that bring about a phenomenon with regularity and without gaps, missing entities, or activities. According to the most influential account of NME (that of Machamer, Darden, and Craver, see [314]), a mechanism is a system of entities and activities organized to produce changes from start to termination conditions. In this view, explaining an event consists in describing the mechanisms that produced it.

Since these accounts of explanation will suffice for our purposes, we will mention some other examples only briefly to draw a more complete picture. In his causal interventionist account, Woodward tries to capture very general features of causal explanation explicitly in terms of counterfactual dependencies [499]. An explanation should answer the question, "what if things had been different?" in terms of what caused the explanandum. This stands in contrast to causal-mechanical accounts of explanation, as both Salmon and Dowe were suspicious of appeals to counterfactual conditions to spell out causality. Finally, Borsboom [78] devised network models of explanation to fit the need in sciences like psychopathology. A network explanation appeals to the topological properties of a network model describing the system in order to explain, for instance, pathological behavior that deviates from the norm.

In all of these approaches, two factors are worth emphasizing. First, most of them assign some role to the specification of causes in explanations, but differ on whether it is more important to specify actual causes or whether good scientific explanations depend primarily on the laws of nature. However, although there is considerable disagreement among philosophers about what (if any) difference there is between causal and non-causal explanations, most agree that many scientific explanations make use of information about causes. [363]

The debate over the importance of causality and laws of nature for scientific explanation continues today, with a growing number of accounts that try to unify both under one umbrella. Against this backdrop, Psillos [381] argues that the concepts of causality, laws of nature, and explanation form a tight web [363]. Likewise, Bartelborth [55] and Overton [364] propose models that try to avoid the problems of previous accounts while harnessing their advantages.

Second, according to all these accounts, what makes something an explanation is whether it accurately depicts the world through the process of explanation pertinent to the account in question [125]. According to DNEs, for example, an explanation is a deductive inference from at least one law of nature and at least one initial condition. So, all of these accounts assume, at least in a certain sense, a specific structure of explanation.

This stands in contrast to most accounts of explanation that are not dedicatedly scientific. For them, it is often the case that a number of pragmatic conditions determine whether something counts as an explanation. One very important pragmatic condition, which will play a role later, is whether the explanation leads to *understanding*. This is not to say that scientific explanations do not value understanding, but rather that producing understanding is just one (although important) epiphenomenon of scientific explanations (see, e.g., [290]) [363].

Finally, there is no such thing as quality criteria for these accounts of scientific explanations. Therefore, it does not matter whether they fulfill a certain set of explanatory virtues or anything like that: as soon as the pertinent form for an account is present, we have an explanation (without any qualification, gradation, or further determination).

8.1.2. Reason Explanation

The above accounts of explanation are all concerned with scientific contexts. Although explanation and science are intimately linked, explanation's relevance extends beyond science. Every day, explanation processes take place between people to answer questions like "Why did you lie to your grandma?" or "Why did you miss our meeting?", and explanations given in response to these questions are presumably not scientific. Nevertheless, there is philosophical interest in these questions. This interest, however, is not from the philosophy of science but from the philosophy of action. Among the many ways to answer questions like these, one way is of particular interest to philosophers of action. In this respect, Alvarez writes:

> "A person's action may be explained in a variety of ways: by reference to the agent's goal, or habits, or character traits, or to her reasons for acting. For instance, we may say that Jess went to the hospital in order to reassure her father, [...] or because her father was in intensive care. These statements explain why Jess went to the hospital because, given certain background assumptions, they enable a third person to understand Jess's action: they make it intelligible. [...] Among this variety of possible explanations (and there are more), the last one is a distinctive type that is of particular interest here because it is an explanation of an intentional action that rationalises the action: it explains the action by citing the agent's reason for acting" [15].

What Alvarez describes here are so-called *reason explanations*. In general, reason explanations are the explanations that humans typically use when trying to understand and explain action, when exchanging justifications for actions and recommendations, and when trying to resolve disagreements [14, 223]. Reason explanations have some history in philosophy, having been first described by Davidson in the early 1960s:

> "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action—some feature, consequence, or aspect of the action the agent wanted, [...], thought dutiful, [...] or agreeable" [141, p. 685].

Let us first clarify what reasons are and which kinds of reasons figure in reason explanations. In the philosophy of action, reasons are categorized by the distinction between normative and motivating reasons [15, 223, 319]. *Normative reasons* are facts that objectively favor or disfavor an action. All normative reasons, taken together, make the action right or wrong. For example, the fact that eating vegetables is healthful counts in favor of one eating vegetables.

Although, ideally, a person has normative reasons available, reason explanations will instead focus on *motivating reasons*, because people can make mistakes. A motivating reason is a consideration that an agent relies on in acting, a consideration "for which someone does something, a reason that, in the agent's eyes, counts in favor of her acting in a certain way" [15]—whether or not it is a fact and actually favors the action.

Motivating reasons stand at the intersection between explanation and justification, as they help to explain an action in the light of what the decider took to justify or favor it [223]. In contrast to normative reasons, motivating reasons can include merely apparent facts, that is, non-obtaining states of affairs or false propositions that the agent mistakenly believes to be true [133, 416]. For instance, that spinach is a good source of iron is a merely apparent fact. Even though it is not the case that spinach is a good source of iron, this can be the reason that motivates a person to eat spinach—since that person mistakenly believes that spinach is a good source of iron, this favors the action in their eyes, and it is the light in which they act. If a motivating reason is not mistaken, we say that it *corresponds* to a normative reason.

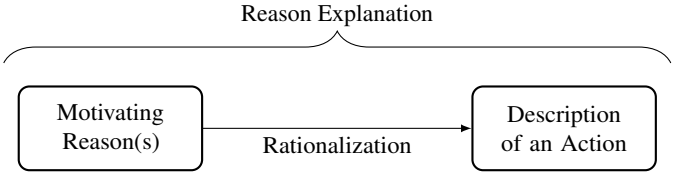


Figure 8: The model of reason explanation.

In reason explanation, an action is explained by the motivating reasons of the agent—that is, the information or misinformation that led him or her to take that action. These reasons *rationalize* the agent's action. Although agents may be aware of numerous pro and contra considerations, and may be led to an action by such a bundle of reasons, most reason explanations of human actions focus on only one or a few contextually relevant motivating reasons. Figure 8 visualizes reason explanations based on the general model of explanation.

8.2. From Explanations to Machine Explanations

There are many other accounts of explanation, and a comprehensive review is beyond the scope of this thesis. These few accounts, however, should suffice to illuminate the concept of explanation, and they will be used later when assessing the appropriateness of our argumentation-based approach. In short, we will argue that, in principle, our approach can be used to generate all of these kinds of explanations (see Section 11.1.1) and, thus, that it harnesses their individual advantages.

8.2.1. Important Distinctions Concerning Explanations

As a step from explanations in general towards machine explanations, we will briefly distinguish between narrow vs. broad views of explanation, and between process and product.

Let us first discuss the distinction between process and product. Nouns ending in the suffix, *-tion*[38] can denote either of an activity or the result of this activity [115], and the term "explanation" suffers the same fate. Achinstein writes the following on this ambiguity:

[38]These Latin-root words originate in verbs, nominalized in the third declension. For example, *ago* I act → *actus* → *actio/actionis*; *explano* I explain → *explanatus* → *explanatio/explanationis*.

> "Suppose that Dr. Smith explained Bill's stomach ache by saying that Bill ate spoiled meat. There is some *act* that has occurred, viz. the doctor's explaining, which went on for some period of time however short. There is also the *product* of this act, viz. the explanation given by the doctor, which did not go on for any period of time but was produced in or by the act of explaining" [2, p. 1].

As Achinstein illustrates, the expression "the explanation given by the doctor" can be used to refer either to the *act* of explaining or to the *product* of this act [363].

Beside the process–product ambiguity of "explanation", there is another distinction that is worth highlighting [363]. Again, a quote from Achinstein serves as illustration:

> "The terms 'explain' and 'explanation' can be used broadly to refer to explaining acts and products that may or may not be good (adequate, successful, 'scientific'). They can also be used more narrowly to refer only to acts and products that are (regarded as) good. In accordance with the broader, but not the narrower, use an atheist could admit that his religious friends are explaining the origin of man when they assert that man was created by God. And he could refer to the product of such acts as an explanation" [3, p. 4].

What Achinstein is outlining here is that there are at least two further uses of "explanation" (primarily in the product sense): a narrow and a broad one [363]. When used broadly, the term "explanation" refers to explanations that may or may not be "good". In the narrow use, the term refers only to the explanations that are judged to be "good". So, in order to understand what explanations are in the narrow sense, the question now arises as to what it means for an explanation to be "good".

With this question in mind, we can begin to approach machine explanations. More specifically, what can be taken to delineate machine explanations from other types of explanations is the criterion for "goodness". As described earlier, a scientific explanation is "good", when it accurately maps onto the world. The common view in the machine explainability community, however, is that an explanation is "good" when it evokes understanding in an addressee (see, e.g., [54, 99, 339, 393, 399]). This brings machine explanations close to reason explanations, where the cited reasons are, ideally, ones that rationalize the action for an addressee.

Employing both of the above distinctions, discussions of machine explainability often take the view that the explanation process is not finished until an addressee has gained an understanding of what was intended to be explained [231, 232]. Let us again emphasize the difference from scientific explanation: here, an explanation is a full-fledged product as soon as it fulfills the form that is pertinent to, for instance, one of the presented accounts.

The next question that arises now is what it is that is to be invoked by machine explanations to facilitate understanding. So far, we have introduced the system's visible behavior, the algorithm on which this behavior is based, and the input needed to produce a certain output as possibilities for explanantia of machine explanations (see Section 1.2 and Section 3.2).

8.2.2. The Explanans of (Machine) Explanations

We will now look in more detail at the possible explanantia for machine explanations and, more generally, the problems associated with determining them. Since these problems are not specific to machine explanations, we will again first describe the general picture and then highlight characteristics particular to machine explanations.

Once more, Aristotle can serve as a good source for discussion. As mentioned above, he noticed that the same "why"-question can be answered in very different ways. In this regard, however, the distinction between *rationes cognoscendi* and *rationes essendi* is only one of two important distinctions, and concerns what he called *reasons* or *principles* (ancient Greek: *archaí*). The second distinction concerns what he called *causes*[39] (ancient Greek: *aitíai*). Aristotle distinguished four such causes, and his *Four Causes* model, also known as the *Modes of Explanation* model, is still influential today. In the second book of *Physics*, Aristotle wrote:

> "Knowledge is the object of our inquiry, and men do not think that they know a thing till they have grasped the 'why' of it [...]. In one sense, then, (1) that out of which a thing comes to be and which persists, is called 'explanation' [...]. In another sense (2) the form or the archetype [...] and its genera are called 'explanations' [...]. Again (3) the primary source of the change or coming to rest [...]. Again (4) in the sense of end or 'that for the sake of which' a thing is done [...]. This then perhaps exhausts the number of ways in which the term 'explanation' is used [...]. As the word has several senses, it follows that there are several explanations of the same thing." [210, Book II, Part 2]).

The bottom line of Aristotle's observation is that there is nothing like a one-size-fits-all explanation of a thing. Rather, the same thing can be explained in different ways, all of which are equally valid. These different ways are nothing more than different aspects of the same thing, each of which can serve as an explanation's explanans:

- **Material:** The substance or material from which something is generated or of which it is made. Material explanations are sometimes referred to as *categorical explanations*.
- **Formal:** The form, structure, or properties of something that make it what it is.
- **Efficient:** The proximal mechanisms or the cause responsible for the change in something. Efficient explanations are sometimes referred to as *mechanistic explanations*.
- **Final:** The what-it-does that makes it what it is; its purpose or what it serves for. Final explanations are sometimes referred to as *functional* or *teleological explanations*.

[39] Aristotle's use of the term "cause" does not refer to what we mean by the term today, nor to what modern theories of causality mean by it. Only his efficient causes are the best fit with the modern use of "cause". [363, 364]

Autonomous Vehicle #9

Let us illustrate the four causes using an autonomous vehicle.

- **Material:** Steel and aluminum are the material causes for the vehicle.
- **Formal:** Having a certain number of tires (e.g., four for a car), an engine, seats, etc. arranged in a specific layout is the formal cause of an autonomous vehicle.
- **Efficient:** A vehicle manufacturer is an efficient cause of an autonomous vehicle.
- **Final:** Transporting people and cargo from one place to another without human intervention is the the final cause of an autonomous vehicle.

According to Aristotle, one must grasp all four causes of a thing in order to fully understand it. This, however, implies that we can gain a partial understanding of a thing by grasping one or more causes of that thing. Indeed, different individuals might be more interested in one Aristotelian cause than another.

Example #16

An artisanal carpenter, for example, might be more interested in the formal and material causes of a chair, whereas a person uninvolved and uninterested in craftsmanship might be interested only in the chair's final cause: that one can sit on it.

The same is true for machine explainability. As we will make visible in Section 9.2.1, depending on the goal that is pursued with explainability, the explanans of interest may change. When it comes to these explanantia, however, a large variety can be observed.

To identify potential explanantia for machine explanations, we reviewed over 200 papers on explainability from the last nearly 40 years and found the following options (see [105, 107]): the system in general (e.g., global aspects of a system) [99], and, more specifically, its reasoning processes (e.g., inference processes for certain problems) [373], its inner logic (e.g., relationships between the inputs and outputs) [277], its model's internals (e.g., parameters and data structures) [236], its intention (e.g., pursued outcome of actions) [234], its behavior (e.g., real-world actions) [190], its decision (e.g., in terms of underlying criteria) [4], its performance (e.g., predictive accuracy) [304], and its knowledge about the user or the world (e.g., user preferences) [190]. Given this multitude, we will continue to talk about "aspects of a system" when it comes to the explanans.

One vital point that these possibilities show is that explanantia in machine explainability are mostly linked to the system itself, rather than to processes outside of it. The aim can be to explain why or how a certain prediction was made, why the system exhibits a certain performance, why the system made a certain mistake, and many more outcomes.

Digression #6

Imagine an artificial computing system that forecasts the weather. This system predicted rain for a particular day and place. However, on the day in question, the weather at the place in question (and also in a wide region around that place) is sunny.

In this case, we might want to know what went wrong with the forecast and ask for an explanation. Obviously, this explanation should *not* be aimed at uncovering the process that led to the weather being sunny at this particular place. Rather, the explanation should aim to uncover one of the system's aspects (e.g., its internal reasoning processes) that led to its prediction of rainy weather.

Optimally, this reasoning process can be rendered intelligible to meteorologists, such that they can check whether the system's reasoning steps correlate with reasoning steps they consider plausible for forecasting the weather at a particular place and time.

By saying that the explanans is rather linked to the system itself, we do not want to exclude that the explanation can also be useful for gaining insights about processes outside of it. In some cases, the explanations that are gained in this manner can be used to make predictions about the world outside the system (more on this in Section 9.2.1).

It is by now clear that the concept of "explanation" is very complex, and a comprehensive presentation would require a monograph of its own. We will conclude this section with one last concept related to explanations, which will become interesting later (in Section 11.1.3).

8.2.3. Levels of Explanation

Several authors have proposed that, when explaining a complex system, one can choose to do so at varying levels of abstraction. These *levels of explanation*, as they are sometimes called, also concern the explanans and differ with respect to the entities that are invoked in it.

Dennett proposed that people can take three increasingly abstract *stances* towards explaining objects: the *physical stance*, the *design stance*, and the *intentional stance* [144]. On the *physical stance*, we are concerned with the physical laws that govern a system. We take the physical stance, for instance, when we explain how an artificial system works based on electrical currents, Ohm's law, and so on. The *design stance* is more abstract than the physical stance, and does not presuppose any knowledge of physical processes. Taking this stance, we explain a system based on its design goal or purpose (similar to Aristotle's final cause). The *intentional stance* is the most abstract stance. Again, this stance does not presuppose any knowledge of processes involved in either of the other stances. In the intentional stance, we ascribe mental states and human intentions to a system, on the basis of which we explain it.

There are other scholars who advocated three-level models, especially in cognitive science. Marr [320], building on earlier work with Poggio [321], introduced three *levels of analysis* for computational problems: the hardware implementation level, the representation and

algorithm level, and the computational theory level. These levels were intended to facilitate the understanding of human neural circuits (particularly for the visual cortex) through an understanding of artificial computation. Similarly, Newell [351, 352] proposed three *levels of description* (physical/device level, program/symbol level, and knowledge level) for the representation of knowledge. To the same end, Pylyshyn [387] distinguished three *levels of organization* (physical/biological level, symbol level, and semantic/knowledge level). [363]

Plausibly, these *levels of explanation* are of particular importance in the field of machine explainability. As emphasized above, the central goal in machine explainability is to evoke understanding in addressees. Obviously, different addressees have different goals and background knowledge. Accordingly, it is likely that they also differ in what is needed to make them understand. Using different levels to explain something is one way to tackle this issue.

Example #17

Let us return to the chair (Example #16). Here, we sketched how an artisanal carpenter and a person uninterested in craftsmanship might differ in their primary interest in a chair. This is because of their different backgrounds and goals. The woodworker, driven by artistic ambition, might want to recreate a particular distinctive specimen of a chair. To this end, at least knowledge about its material is required. The uninterested user, however, will in most cases not care about a chair's material, as long as one can sit down on it.

Now, a similar divergence can be observed with respect to Dennett's stances, although the chair example does not work well here. So, let us return to our autonomous vehicle example:

Autonomous Vehicle #10

Imagine that an autonomous vehicle causes a serious accident by crashing into a wall. Now, several parties are involved in finding out the cause of this accident.

Hardware experts could approach this task, for instance, by checking the vehicle's LiDAR using complex optical laws to find out whether it could detect the wall at all. Accordingly, they take the physical stance towards explaining the accident.

Software experts, on the other hand, could approach this task by checking the vehicle's AI to see whether the collision-avoidance routines were working properly. Here, one could argue that they take the design stance towards explaining the accident.

Finally, *insurance agents* might not be directly involved with the vehicle, but rely on abstract reports by the hardware and software experts. In such reports, a misclassification by the vehicle AI could be reframed so that the report states, for example, that the vehicle thought the wall was heavy rain and, therefore, continued to drive to reach its destination. In essence, such a description attributes human-like traits to the vehicle, thereby adopting the intentional stance.

This example illustrates that, based on their background, different addressees are likely to need explanation on different levels. An insurance agent or a software expert will usually not be able to understand the intricacies of optical theory that led the vehicle to crash. Likewise, the insurance agent will most likely also not be able to understand certain code decisions that led the vehicle's algorithm deviate from its intended purpose.

This brings us back to the different types of explanation we introduced. To recall, a DNE makes reference to physical laws and a CME to causal processes at play. For this reason, these explanation accounts seem to be well-suited to give explanations conforming to the physical stance. NMEs, on the other hand, seem to be better suited for explanations with respect to the design stance. Finally, reason explanations are the perfect fit for the intentional stance, as they refer to the beliefs and motivations of an agent.

8.3. What is Machine Explainability?

Now that we have a rough understanding of explanation per se and of machine explanations, we can turn to *machine explainability*. As the term implies, machine explainability is the ability to generate explanations of (certain aspects of) a machine. We will discuss this concept in more detail later, after a short primer on machine explainability.

8.3.1. A Short Primer on Machine Explainability

Research on ways to explain and justify how artificial systems work has been going on since at least the 1970s.[40] While the goal at that time was to explain rule-based expert systems, the focus later expanded to other types of systems. In the context of recommender systems, for instance, the interest in explaining systems intensified in the early 2000s (see, e.g., [222, 463]).[41] However, it was not until the advent of powerful ML techniques such as deep learning (DL) that research on machine explainability attracted interest from a broader audience.[42] We will take a look at possible reasons for this development in Section 9.2.1.

Throughout these years, however, research in machine explainability has remained largely uninfluenced by common (philosophical) theories of explanation. Instead of using theories devised by philosophers or other scholars that are professionally involved in researching explanations (e.g., psychologists and other social scientists), machine explainability researchers have mostly developed their own approaches, and many publications contain the authors' own ideas of what explanations are [340].

[40] The earliest research paper on this topic that we found dates to 1975 (namely, [428]). The most famous project in this early research phase (described in [91]) is to explain the expert system MYCIN.

[41] While we think that some notions of causation and accountability in computer science (not to be confused with their legal or philosophical counterparts) can be seen as approaches to machine explanations (see, for instance, [207, 208]), this must be explored in some other location since it is not relevant for our purposes. [446]

[42] A Google Trends search for the terms "explainable" and "explainability" revealed a constant, but low, interest from the earliest accessible date, with a rapidly increasing interest starting in 2017 (see Figure 32).

Although some researchers have begun to call for more interdisciplinarity, even referring to the current trend as "inmates running the asylum" [340], interdisciplinary approaches are only spreading slowly across the research landscape. The lack of a common ground and the sudden surge of interest in the topic have made machine explainability a vast field of research in which it is difficult to keep track of proposed approaches and terms used [447].

Complicating matters further is the fact that research on the explanation of artificial systems is not uniformly labeled. There are several competing terms and no consensus on their meaning. With the term "machine explainability", we aim to create an umbrella category for these related terms. However, to avoid contributing to confusion in the landscape of explainability research by introducing a new term, we will delineate some of the dominant terminology below. This should also clarify what we conceive machine explainability to be.

8.3.2. Delineating Machine Explainability

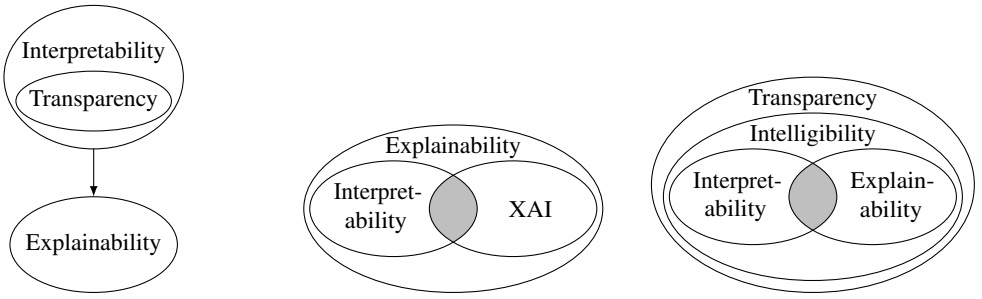
The literature mentions, among other terms, "interpretability", "explainability", "explainable artificial intelligence", "perspicuity", "explicability", "transparency", and "intelligibility", sometimes synonymously, sometimes not. A good overview of the number of terms related to explainability is provided by Vilone and Longo, who collected 36 of them with their associated descriptions as part of a literature review [480].

What connects these, at first sight, very different terms is the implicit assumption that the transparency/intelligibility/perspicuity/etc. of a system is brought about by explanations. However, this is where the similarity ends. Many authors have attempted to explicate one or more of these notions, with contradicting results (see [102, 116, 399]), and a comprehensive review of all these notions would warrant a work of its own. Nevertheless, we will address two of the most important terms in what follows.

We will first briefly discuss *explainable artificial intelligence (XAI)*. XAI is currently a very prominent research area [54], concerned primarily with explaining (certain aspects of) systems based on *AI*. However, while the opacity of many AI-based systems is one of the driving forces behind the renewed interest in making systems explainable, even (aspects of) systems that are not based on AI can exhibit a high degree of opacity. For this reason, we believe that the scope of XAI is too narrow. Accordingly, a more general focus on machine explainability, the goal of which is to make *all* kinds of artificial computing systems explainable, is more beneficial in our eyes.

This should suffice as a comment on XAI. There is also another term that is important to mention. In research concerning ML, the term "interpretability" is more common than the term "explainability". Although there is research trying to differentiate the use of these two terms (e.g., [116, 263, 399]), the results are inconsistent, sometimes indicating that "interpretability" is the more general term, sometimes indicating the same for "explainability".

For illustration, we have extracted some visualizations from the literature that aim to show the relationship between (at least) these two terms (see Figure 9). As can be seen, Figure 9a



(a) Interpretability being a precondition for explainability [399].

(b) Interpretability being a subconcept of explainability [263].

(c) Interpretability and explainability having intersecting fields of application [116].

Figure 9: Different relationships between explainability and interpretability.

depicts interpretability as a prerequisite for explainability. Figure 9c, on the other hand, depicts that explainability and interpretability only intersect. To better understand machine explainability, it is beneficial to know how it differs from interpretability. To this end, we will have a look at some quotes in what follows:

> "As we explain in this section, a system's level of explainability is created through the interpretation the agent provides" [399, p. 674]. (1)

> "To *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts. Therefore, in data mining and machine learning *interpretability* is defined as the ability to explain or to provide the meaning in understandable terms to human" [201, p. 5] (see also [155, p. 2]). (2)

> "Interpretability is about the extent to which a cause and effect can be observed within a system. Or, to put it another way, it is the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters. [...] Explainability, meanwhile, is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms" [183]. (3)

> "Systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation." [73, p. 8]. (4)

> "Although interpretability and explainability have been used interchangeably, we argue there are important reasons to distinguish between them. Explainable models are interpretable by default, but the reverse is not always true" [188, p. 80]. (5)

> "We argue that interpretation is a relation between two explanations. During the process of interpretation, one explanation gives rise to a *more understandable* explanation" [171, p. 835]. (6)

> "I equate interpretability with explainability" [339, p. 8]. (7)

> "'Explainability' is preferred here to 'interpretability' to highlight that the explanation of a decision must be comprehensible not only to data scientists or controllers, but to the lay data subjects (or some proxy) affected by the decision" [342, p. 17]. (8)

> "Interpretability enables transparent AI models to be readily understood by users of all experience levels. Explainable AI applied to black box models means that data scientists and technical developers can provide an explanation as to why models behave the way they do – and can pass the interpretation down to users" [299]. (9)

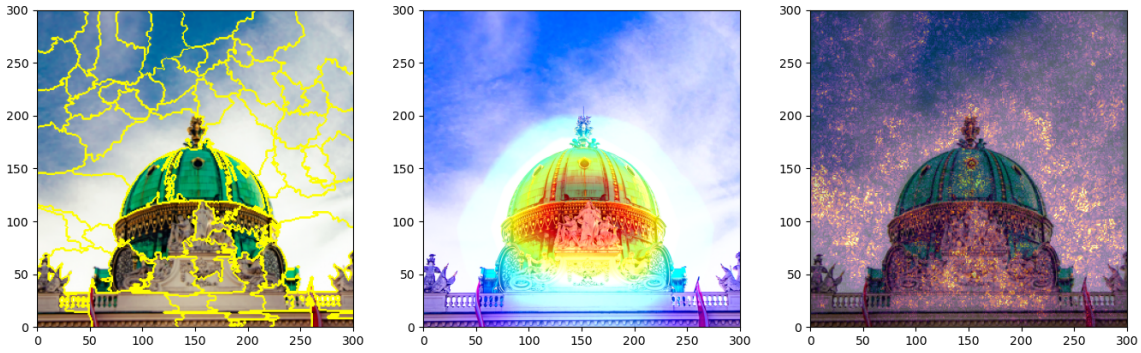
These examples should serve as a rough indication of the confusion that pervades the literature. However, this confusion shall not be to our disadvantage. In contrast, it shall serve as a stepping stone for us, as conceiving a clear distinction between interpretability and explainability will allow us to better elaborate our understanding of machine explainability. Of course, our proposed distinction will not adequately account for all uses in the literature, as some of them are contradictory (see, e.g., Quote 8 versus Quote 9). Nevertheless, our distinction will provide a basis for understanding a large part of the literature.

Before discussing the individual terms, let us first highlight a common feature of these quotations, namely, their emphasis on human *understanding* (see, e.g., Quote 2, Quote 4, Quote 6, Quote 8, and Quote 9). Indeed, these quotations confirm our earlier claim that the common expectation in research on explanations of artificial systems is that these explanations increase a human's understanding of (certain aspects of) a system.

With this in mind, let us now turn to the distinction between the terms. We assume that the goal of much research on interpretability is to map human comprehensible concepts to the processes occurring in artificial systems. Research on explainability, on the other hand, is concerned with communicating these concepts to different kinds of human recipients, so as to evoke their understanding of the original processes. Let us elaborate.

What makes artificial systems so difficult for humans to understand is their way of reasoning. This can easily be illustrated by examples. While it is possible for an expert to make sense of the reasoning processes underlying knowledge-based systems, this is not necessarily the case for laypersons. Furthermore, when it comes to systems using ANNs, even experts have difficulties making sense of their reasoning processes. In the latter case, this is at least partly due to their sub-symbolic nature. While knowledge-based systems are designed to work with a representation of knowledge that is human interpretable, this is not the case for ANNs.

Research on *interpretability* aims to amend this problem. More specifically, this research aims at devising methods and techniques to map human-understandable concepts to the sub-symbolic processes in a system. Without an interpretation, we may only see certain activation patterns of neurons in ANNs. With an interpretation, we can at least map the activation of certain clusters of neurons to, say, the identification of curves in a picture.



(a) Segmenting into interpretable superpixels.

(b) Highlighting by using a heatmap.

(c) Highlighting by showing salient pixels.

Figure 10: Different techniques to assign an interpretation to a picture.

Figure 10 shows the approaches of several interpretability techniques. Note that, because many ML-based systems contain sub-symbolic processes that are difficult for a human to interpret, research on interpretability can be found predominately in the ML community.

Understood in this way, interpretability is a prerequisite for explainability. Linked to our previous findings that if the result of an action is increased understanding, then the action qualifies as an explanation, we can infer that only the use of humanly comprehensible terms while explaining a system may succeed in increasing explainability (see also Section 10.2.2).

Interpretability, however, is just one step in understanding a system. In many cases, the interpretations obtained using interpretability techniques are hard to parse for laypersons and can only improve the understanding of experts (see, e.g., Figure 10c). Against this background, one of the goals of explainability is to develop methods that further enhance such interpretations such that they are also useful for increasing the understanding of non-experts.

Let us check our conceptualization of these two terms against the preceding figures and quotations. Overall, a concept similar to ours is directly reflected in Figure 9a and in Quote 1, Quote 5, and Quote 8. All these quotes can be construed in the sense that interpretability is a prerequisite for explainability. Furthermore, the concept can be found indirectly in Figure 9b and in Quote 2. In this quote, there is simply no reference to explainability that would allow for a direct link. However, the description of interpretability is compatible with our concept. Thus, most of the figures and quotations we extracted are consistent with our proposal.

However, this does not explain the other figures and quotations that make statements about how interpretability and explainability are related. We conclude this discussion by briefly commenting on these quotations. First, we can disregard Quote 7, since it does not purport to make a distinction. Furthermore, while Quote 4 seems to make a valuable distinction, this distinction seems to be shifted by a level: “explainability” is used to denote what we call “interpretability”, and “interpretability” is used to denote a step even prior to what we call “interpretability”. Accordingly, our distinction can also be found in this quote, only under a different name. We will come back to Quote 6 and Quote 9 later.

8.4. What to Expect from Machine Explainability

At this point, we have worked out that the explainability of systems should contribute to their understandability. But what does this understanding bring us? We argue that there are further, downstream, goals of explainability. Indeed, evoking understanding of (certain aspects of) artificial systems is not the ultimate goal of machine explainability. We claim that this understanding is in most cases only an intermediate step to achieve other goals: the *desiderata*.

In Section 3.2, we have already outlined some desiderata. For example, a greater understanding of a system should help attribute responsibility, or it should increase the system's maintainability. Overall, we have argued that machine explainability is essential to reap the advantages associated with machine ethics while avoiding its drawbacks. There are more desiderata than we have discussed so far, and we will review them in detail in Section 9.2.1.

Generally, however, desiderata arise from people who have interests, goals, expectations, needs, and demands regarding artificial systems. There are many groups of such people, with various interests. For example, various people operate the systems, try to improve them, are affected by decisions based on their outputs, deploy the systems for everyday tasks, and set the legal framework for their use. These people are commonly grouped as *stakeholders* (e.g., users, developers, etc.).[43]

Basically, the interests of these stakeholders (e.g., to have fair or trustworthy systems, see [155, 172]) are nothing other than the desiderata. Generally, many desiderata call for greater understandability of artificial systems, and it is assumed (as elaborated so far) that explainability can lead to this understandability by providing individuals with the explanations required to understand (certain aspects of) artificial systems [54, 105, 294, 490].

8.4.1. A Model of Machine Explainability

This process can also be presented in a more orderly and articulated way. To this end, we present a conceptual model (see Figure 11) that organizes and makes explicit the central concepts in machine explainability and their relations, and how they relate to satisfying the desiderata. The main concepts in this model are: "explainability approach", "explanatory information", "stakeholders' understanding", "desiderata satisfaction", and "(given) context".

At this point, however, two components of our model are still missing: explainability approach and context. We introduce these components first, and then turn to the model itself.

The first missing component is the approaches that enable or provide insights into (certain aspects of) artificial systems. These approaches (we call them *"explainability approaches"*) encompass methods, procedures, and strategies for providing explanatory information that help someone to understand artificial systems better [294]. A specific explainability approach is characterized by all the steps and efforts taken to extract explanatory information from a system, and to provide it adequately to an addressee in a given context [294].

[43] A stakeholder is, among others, someone "who is involved in or affected by a course of action" [336]. We use this as a general term, but refer to specific stakeholder classes where appropriate.

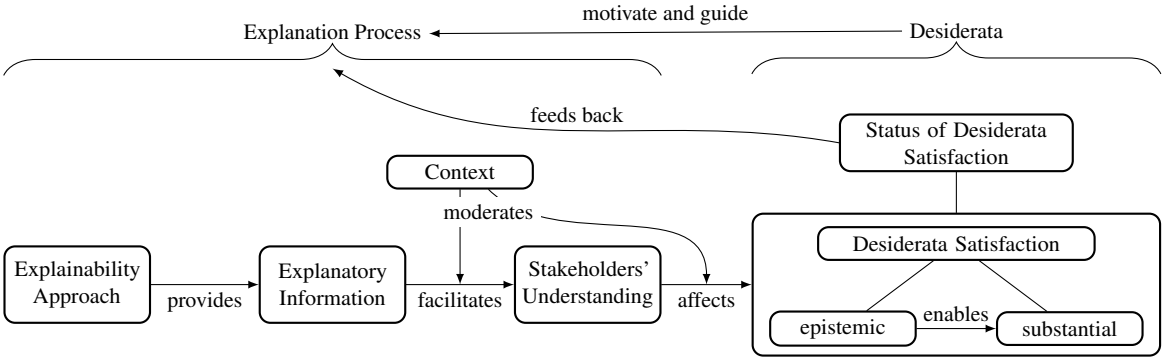


Figure 11: Our proposed model of the main processes in machine explainability.

We need to make two important comments here. First, by introducing the term "explainability approach", we want to avoid the confusion between interpretability and explainability by effectively ignoring this distinction in what follows. In other words, an explainability approach may or may not include interpretability techniques and methods. Second, by speaking of "explanatory information", we take into account that, in machine explainability research, a set of propositions is not an explanation until it increases understanding.

The other missing component is the *context*. There is no agreed-upon definition of the term "context" [64] (for discussions on this topic, see [64, 157]). Following Dourish [157], we hold that context is determined by a given situation, in the interaction between a stakeholder, an artificial system, a particular activity or task, and an environment. Without knowing the specific situation, anticipating all the contextual influences that will affect the process of how explainability approaches should satisfy desiderata is impossible.

The basic idea of our model is that the success of an explainability approach depends on the satisfaction of desiderata (consisting of the *substantial* and the *epistemic* facets of desiderata satisfaction; see below). Desiderata satisfaction, thus, motivates an explanation process including explainability approaches, explanatory information, and understanding.

In the explanation process, we assume that explainability approaches provide explanatory information to human stakeholders. These stakeholders engage with the information to facilitate their understanding of (certain aspects of) an artificial system. As a result, the stakeholders' adjusted understanding affects the extent to which their desiderata are satisfied. The context in which the stakeholder and the artificial system act and interact affects the relations between the other concepts (i.e., it affects the relation between explanatory information and understanding, as well as the relation between understanding and desiderata satisfaction).

The upcoming sections are guided by this model, where we will begin by walking backwards through the model and discuss desiderata. After that, we turn to explainability approaches, survey some exemplary cases, and discuss what makes a good one. However, we will not address the model's intermediate steps, such as understanding and context. These are open research topics that warrant (and have warranted) separate work of their own.

Before we can address desiderata individually, however, we must first clarify, in general terms, what it means to fulfill a desideratum. To this end, we provide some practical examples.

8.4.2. Desiderata Satisfaction

The stakeholders involved in explainability are diverse. Accordingly, the motivations for acquiring explanations about (certain aspects of) artificial systems are at least equally diverse. As the need for explainability arises when desiderata are not (sufficiently) satisfied [54, 173, 175, 341, 379], we have to clarify what it means for a desideratum to be satisfied.

We propose that the satisfaction of each desideratum comprises two facets. We call these facets *epistemic* and *substantial* desiderata satisfaction, respectively. On the one hand, stakeholders want systems *to have* certain properties that actually make them fair, transparent, or usable. In line with this, the substantial facet of a desideratum (e.g., fairness) is satisfied if a system sufficiently possesses the corresponding properties (e.g., if the system is, *de facto*, fair).

On the other hand, stakeholders want *to know* or be able *to assess* whether a system (substantially) satisfies a particular desideratum (i.e., whether the system has the desired properties). Thus, the epistemic facet of the fairness desideratum is satisfied for a stakeholder if they are in a position to assess or know whether and to what extent the system is fair.

Example #18

Take the desideratum of having usable systems. A successful explanation process, as depicted in our model, can enable users to recognize whether a system is usable, and, optimally, also increase the system's usability to a certain degree. In this case, the epistemic satisfaction consists in the stakeholders being able to assess whether a system or its outputs are usable for the task at hand. To a lesser extent, however, an explanation process can also contribute to the substantial satisfaction of the desideratum, since it provides additional knowledge about the system that makes it more usable for the stakeholder. For larger deficits in usability to be addressed, however, explanatory information might not directly help; for this, the entire system may need to be redesigned.

Depending on the desideratum, the two facets are correlated to a certain degree (possibly even completely if satisfying the epistemic facet to a certain degree satisfies the substantial facet to the same degree).

Example #19

Consider the desideratum of retaining user autonomy in human-in-the-loop scenarios. Suppose that an explanation process has helped to satisfy the epistemic facet of this desideratum to some extent, since it has enabled the user to assess the extent to which they can retain their autonomy in making decisions based on the system's recommendations. Additionally, the better a user understands the output of a system, the more autonomously they can decide based on it. Thus, the explanation process has helped sat-

isfy both the epistemic facet and the substantial facet of this desideratum. Accordingly, in this case, the two facets of desiderata satisfaction are highly correlated.

Example #20

Let us consider the desideratum that systems adhere to certain moral principles.[a] When they have sufficient information about a system, regulators can evaluate whether this system meets moral standards. Again, the explanation process serves to satisfy the epistemic facet of this desideratum. However, this does not directly make the system's processes and outputs comply with moral standards. Consequently, explanation processes can at best indirectly satisfy the substantial facet of this desideratum: based on the understanding gained through the explanation process, faults can be identified, and steps taken to improve systems with respect to their moral properties. In this case, the epistemic and the substantial facet of desiderata satisfaction are only loosely correlated.

[a] We ignore here all considerations regarding a (moral) right to explanation [483].

Distinguishing between these two facets is important to highlight that the explanation processes can contribute to satisfying all epistemic facets of desiderata concerning artificial systems. This is also the reason why the epistemic facet is of particular interest for machine explainability: an explanation process alone is sometimes not sufficient to satisfy the substantial facet of desiderata concerning artificial systems.

In many cases, however, the epistemic satisfaction enables the substantial one. That is, a better understanding of systems, while not always directly leading to a substantial satisfaction of desiderata, can provide the necessary foundation for it. Since epistemic satisfaction of a desideratum is closely related to understanding a system better, understanding crystallizes as the linchpin for all efforts to satisfy desiderata. With this knowledge, we can now move on to the next section, where desiderata play a pivotal role.

9. From Machine Explainability to Machine Ethics

In the first part of this thesis, we argued that the need for machine explainability is a by-product of the need for machine ethics. Indeed, with the increasing societal significance of intelligent systems and other circumstances, such as the fact that many of them still need to be operated by humans, these systems are affecting people's lives more and more, which requires a more comprehensive consideration of human well-being and other human interests.

Machine explainability thus links back to machine ethics: Many desiderata that can be satisfied by machine explainability are also related to machine ethics, such as more acceptance and fairness of systems, or the possibility to assign responsibility in cases of failure. Furthermore, the satisfaction of legitimate (i.e., philosophically, legally, etc. justifiable) desiderata can improve the lives of many people, which constitutes another link to machine ethics.

For this reason, and because stakeholders, in combination with their desiderata, motivate, guide, and affect the explanation process (see Figure 11), identifying and clarifying the desiderata and the different classes of stakeholders in the context of artificial systems is crucial to fully explore the benefits of machine explainability. We will do so below by starting with the stakeholders. Afterwards, we will come to the desiderata, and, for one desideratum in particular, we will explore how, exactly, explainability contributes to its satisfaction.

9.1. Classes of Stakeholders

In previous research, varying classes of stakeholders have been discussed in the context of explainability. For example, Preece et al. distinguish between four main classes of stakeholders: *developers*, *theorists*, *ethicists*, and *users* [379].

Arrieta et al. categorize the main classes of stakeholders as *domain experts/users*, *data scientists/developers/product owners*, *users affected by model decisions*, *managers/executive board members*, and *regulatory entities* [54]. Other researchers have distinguished similar classes of stakeholders (see, e.g., [173, 225, 490]). We follow these researchers and distinguish five classes of stakeholders: *(end) users*, *(system) developers*, *affected parties*, *deployers*, and *regulators* (see Figure 12 for a visualization of the stakeholders and their relationships).

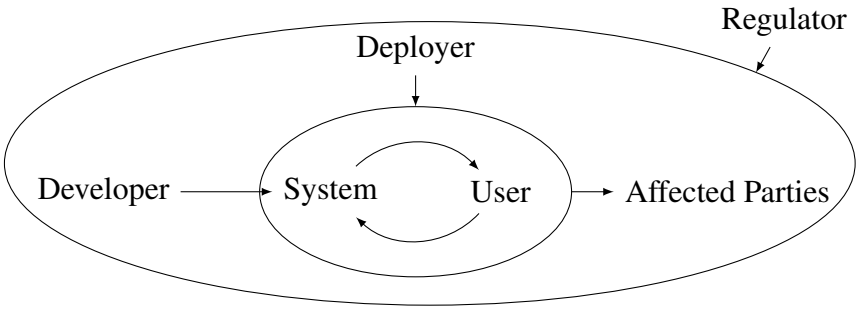


Figure 12: The classes of stakeholders associated with artificial systems and their relations.

Clearly, one person can be a member of multiple stakeholder classes. A user, for instance, may be affected by decisions made using the system they operate. Additionally, these are only prototypical classes of stakeholders, and more finely grained distinctions into sub-classes of stakeholders are possible [173]. For example, there is not one prototypical developer; developers differ in their expertise and in other factors (e.g., personality). A novice developer may desire different things compared to an expert. In a similar way, a lay user's desiderata might differ from those of an expert user. Moreover, we would like to emphasize that this list of stakeholders is not necessarily exhaustive, since our distinction among stakeholder classes is based on previous research that mainly has a computer science background, and might, therefore, neglect other classes of stakeholders. Nevertheless, let us take a look at the proposed classes.

Users Most works on stakeholders in machine explainability have this class of stakeholders in common (see, e.g., [54, 379, 490]). Among others, users take recommendations of artificial systems into account to make decisions [225]. Some prototypical members of this stakeholder class are medical doctors, loan officers, judges, or human resource managers.

In general, users are not experts in the technical details and workings of the systems they use. However, users can work most efficiently if they have reasonable expectations about how a system will work. When users do not have reasonable expectations, or in cases where their expectations are not met, they need information that goes beyond the knowledge needed to operate the system.

Developers The individuals who design, program, and create systems are the developers. Naturally, they count as a class of stakeholders because, without them, the systems (or the programs running on them) would not exist in the first place. For this reason, developers are also frequently found in papers on stakeholders in explainability (see, e.g., [54, 379, 490]).

In general, developers have a high level of expertise concerning systems. They need to create systems that work reliably and correctly, and, in the case of systems that do not meet these requirements, they sometimes need to improve them. To this end, they require as much information about the system as possible, especially when it is a system that they did not create themselves. For AI-based systems, they need ways to get behind the opaque workings of ML models.

Affected Parties The influence of artificial systems is constantly growing, and decisions about people are increasingly automated—often without their knowledge. Affected parties are such (groups of) people in the scope of a system's impact. Obviously, they are stakeholders, as, for them, much hinges on the decision of a system. Patients, job or credit applicants, or defendants in court are typical examples of this class.

Affected parties can be from all kinds of backgrounds with all kinds of expertise in artificial systems. What they have in common is that they have been affected—often negatively—by a decision that was based, at least in part, on the outputs of an artificial system. These people want information on how to get a more desirable outcome next time, so as to be not negatively affected anymore, or they want information to assess whether the decision was correct, or at least well justified, in the first place.

Deployers People who decide where to employ certain systems (e.g, a hospital manger decides to bring a special kind of diagnosis system into use in their hospital) are the deployers. We count them as another separate class of stakeholders because their decisions influence many other classes of stakeholders. For example, users have to work with the deployed systems and, consequently, new people fall inside of the range of affected parties.

Deployers, like the users, are usually not experts in artificial systems. Given the high level of responsibility they bear, however, they need to be able to ascertain that the systems they use do not have any negative consequences for individual users or the company as a whole. In addition, they need to know about the performance of the systems in order to judge whether these are fulfilling their purpose and their use is, thus, justified.

Regulators Finally, there are regulators who stipulate legal and ethical norms for the general use, deployment, and development of systems. This class of stakeholders occupies a somewhat extraordinary role in that regulators have a "watchdog" function not only with respect to the systems, but to the entire interaction process between the system and the other stakeholder classes. This class includes ethicists, lawyers, and politicians, who must have the expertise to assess, control, and regulate the entire process of using artificial systems.

Regulators, again, are often people who do not know much about the particulars of artificial systems, and the know-how needed to regulate such systems is usually acquired by consulting experts. Members of this stakeholder class primarily require information about the general functioning of systems. Furthermore, they are also interested in the provision of information about systems to other stakeholder classes in order to preserve their general rights.

Given this diversity of stakeholders, it is clear that there is a wide variety of desiderata. In particular, given the different interests and backgrounds of the stakeholder classes, we want to improve the understanding of desiderata in general. Since desiderata are nothing more than the goals and interests of stakeholders, we need to be aware of their idiosyncrasies when we discuss desiderata below.

9.2. Motivations and Risks of Machine Explainability

In the context of machine explainability, we have so far only spoken of desiderata in a positive sense. However, the desiderata arising from the five classes of stakeholders are numerous and diverse. In particular, stakeholders can also have desiderata that are undermined by machine explainability—we will speak of *desiderata frustration* in such cases.

Here, we consider the desiderata, whose satisfaction or frustration is purportedly influenced by explainability, extracted from a survey of more than 200 peer-reviewed journal and conference publications on machine explainability (see Table 5; see also [105, 106]).[44]

The fact that explainability can also frustrate desiderata gives a reason against making systems explainable. There are also other reasons that speak against doing so. As with machine ethics, then, we must consider whether the advantages of machine explainability outweigh its possible disadvantages. To this end, we will examine the reasons both for and against pursuing the field of machine explainability.

9.2.1. Reasons in Favor of Machine Explainability

We will first start with an extensive discussion about the desiderata whose satisfaction is claimed to be supported by machine explainability. In particular, we will outline each desideratum, depict how its satisfaction can achieved through explainability, and, whenever possible, link it back to the motivations and risks of machine ethics.

Acceptance[45] Deployers want the systems they bring into use to be *accepted*. In the eyes of deployers, the worst-case scenario, in terms of acceptance, is that users reject appropriately working systems, resulting in the systems never being used [477]. Therefore, low acceptance undermines what deployers intend to achieve when providing systems to users.

Previous research claims that machine explainability can aid in this case by investigating how to provide people with more insights into, and understanding of systems, which can increase acceptance [201, 306, 393] (see also Section 3.2.2). In particular, a system that is (perceived as) *trustworthy* can gain acceptance [190], and explainability is key to this.

The acceptance of artificial systems is also a primary motivation for machine ethics (Motivation 1.1). Given the potentially enormous benefits that artificial systems can bring to our society, it is imperative that these systems (or at least some of them) be accepted. Both machine ethics and machine explainability can help in this endeavor, augmenting each other.

Accountability[46] Roughly speaking, *accountability* is the legal counterpart to the more philosophical concept of *responsibility*. Researchers also refer to *liability* (e.g., [345]), *culpability* (e.g., [95]), or *legal accountability* (e.g., [72]).

[44] More details on this literature review and its results can be found in Appendix C. Appendix C contains all quotes from which we have extracted desiderata, broken down by desideratum (Table 6–Table 39).

[45] Representative sources are [73, 188, 201, 306] (see Table 6 for all sources).

[46] Representative sources are [342, 365] (see Table 7 for all sources).

In general, explainability can be used to provide information that allows for entities to be held accountable for a particular outcome [345]. For more details, see the paragraph on *responsibility* below, or our thorough discussion on responsibility in Section 9.3.

Accuracy[47] Developers have a vested interest in a high *accuracy* of the systems they develop. Accuracy is perhaps the most important *performance* measure for systems; it indicates how close the actual result is to the envisioned one.

Although there are some claims that explainability and accuracy are difficult to combine (see, e.g., [366, 399]), there is also the opposite view (see [126] for a discussion). In particular, explainability is seen as a way to help developers estimate the accuracy of systems, and even as a way to actually make systems more accurate (see, e.g., [280, 508]).

In the ML domain, the accuracy of models can benefit from explainability through model optimization [322]. That is, by obtaining information of what leads to the results of an ML-based system, developers can identify underrepresented or erroneous training data and, using this, fine-tune the training process to achieve higher accuracy and correct biases.

Autonomy[48] In a broader sense, preserving a person's *autonomy* means that they have enough information to make a justified decision between available courses of action [175]. If users in the role of decision makers have too little information about how a system arrives at its recommendations, they will hardly be able to make their own justified decisions based on these recommendations; moreover, they will likely have problems explaining why they followed or rejected a system's recommendations (see also Section 9.3.3).

Machine explainability provides adequate means to let users preserve their autonomy when interacting with intelligent systems by providing mechanisms for users to obtain such information [175]. With information provided by explainability approaches, users are empowered both to make justified decisions based on the outputs of artificial systems, and to explain why they decided as they did.

The potential of machine explainability to let humans retain their autonomy when interacting with artificial systems alleviates one risk of machine ethics, namely, that machine ethics might undermine human agency (Risk 2.2.4). Together with machine explainability, human agency can be preserved by enabling people to make autonomous decisions.

Confidence[49] Typically, deployers want the systems they deploy to be used a lot, and appropriately. In addition to *trust* and *reliance*, the users' confidence in a system is an essential prerequisite for this. If users have adequate confidence in a system, they will perform better when working with it than if they have too little or too much confidence.

Confidence is similar to *trust* and *reliance*, and these terms are sometimes used synonymously in the literature. Roughly said, what differentiates these concepts is the type of

[47] Representative sources are [73, 137, 153, 280, 322, 389, 399, 478, 508] (see Table 8 for all sources).
[48] Representative sources are [83, 175, 246] (see Table 9 for all sources).
[49] Representative sources are [30, 51, 54, 73, 274, 348, 373, 508, 510] (see Table 10 for all sources).

justification that one has in using a system. When someone is *confident* that a system works well, they have found some kind of *assurance* that the system does so. In contrast, when one *trusts* a system, they merely *believe* that the system works well without actual proof. Finally, *reliance* does *not* require any form of *justification*. In other words, one would use the system without having any evidence that it works well or without believing that it does so.

Explainability can support the users' confidence in a system by allowing them to understand how its reasoning mechanism works [30], while also preventing overconfidence by highlighting the system's shortcomings. For more details, see the paragraph on *trust*.

Controllability[50] Deployers and users alike wish for the high controllability of artificial systems. In particular, this includes being *educated* about the workings of the system, as well as having the ability to detect and correct errors (i.e., *debuggability*).

Explainability contributes to controllability by helping to *educate* about and *debug* systems (for more information, see the paragraphs on *education* and *debuggability* below). In addition, explainability can also give users a greater sense of control by allowing them to understand the reasons for decisions, and hence to decide whether or not to accept an output [399].

The potential to increase the controllability of artificial systems also aids in mitigating the risk of machine ethics to undermine human agency (Risk 2.2.4). By being able to exert more control over systems, being able to decide whether to accept their outputs or not, and by being educated about their workings, people can retain and sometimes even enhance their agency and autonomy when interacting with systems.

Debuggability[51] One of the most important desiderata for developers is *debuggability*. Modern artificial systems, especially those based on AI, are increasingly hard to maintain. Not only are the computer programs becoming larger and more complex, but the interaction between systems is accelerating, making the identification and rectification of bugs a complicated affair. Accordingly, developers crave ways to improve debuggability, that is, the identification and correction of errors in systems.

Explainability has a positive influence on *debugging* because additional information about a system can help developers identify and fix bugs [4]. Indeed, *debuggability* and *verification* of systems motivated the first efforts for explainability in the 1980s. Specifically, in the case of ML applications, explainability can enable developers to identify and fix biases in the learned model (e.g., by making the training set more diverse), thereby increasing its accuracy (and, hence, its performance) through *model optimization*.

Greater debuggability is essential for making systems less corruptible, thereby alleviating Risk 2.1.1 of machine ethics (increased corruptability). With a higher debuggability, (maliciously) introduced bugs can be better identified and fixed. Overall, the system can be better checked for problematic behavior.

[50] Representative sources are [1, 4, 30, 222, 339, 342, 358, 397, 399] (see Table 11 for all sources).
[51] Representative sources are [4, 73, 131, 137, 341, 342, 391, 399, 497] (see Table 12 for all sources).

Education[52] Education is a very simple desideratum of users, which involves being educated on how to use a system or its outputs.

Explainability can have a positive impact on *education* by providing information that allows users to learn how a system works or how to use it or its outputs competently [137].

Effectiveness[53] Effectiveness has both a system side and a user side, of interest to different stakeholder classes. While deployers are interested in both sides, developers are more interested in the system side. From a system's point of view, effectiveness is a *performance* measure that indicates how successfully a system is in achieving the results it is supposed to achieve. The user side works according to the same principle, and is concerned with how well a user comes to the envisioned results with the support of the system [466].

On the system side, explainability has a positive impact on system *effectiveness* since it has a positive impact on system *performance* (see paragraph below). On the user side, an explainable system can lead to greater decision accuracy by helping users to understand more about a recommended option or product [137].

Efficiency[54] Just as with *effectiveness*, *efficiency* has both a system side and a user side. On the system side, efficiency is another *performance* measure; it concerns the resources (e.g., time, energy) required to arrive at a result. The user side is, again, very similar, indicating the resources that a user needs to come to a decision, using (the output of) a system as an aid.

Similar to the other performance measures (see paragraph below), efficiency can benefit from explainability. However, the gain is potentially less pronounced because making a system explainable may require additional computational resources. Likewise, analyzing and understanding explanations takes users' time and effort [284], possibly reducing their efficiency. Overall, however, the time needed to make judgments could also be reduced by receiving supplementary information [466], thereby increasing user efficiency.

Fairness[55] Considerations of fairness in the context of artificial systems have evolved, at least in part, because, as artificial systems have become more prevalent in our everyday lives, there are ever more affected parties. One hope in introducing automated decision-making processes was that decisions would be less susceptible to human bias [384]. However, it is generally acknowledged that artificial systems can reproduce and, in this process, even reinforce human biases (see, e.g., [97, 303]; see also Digression #1).

Such bias can lead to discrimination against individuals (e.g., in the distribution of jobs, credit, or healthcare), not on the basis of their own actions or characteristics, but on the basis of actions or characteristics of social groups to which they belong (e.g., women, ethnic minorities, etc.) [303]. Therefore, to counteract bias, it is crucial to enable its recognition.

[52] Representative sources are [30, 111, 137, 199, 222, 358, 444, 492] (see Table 13 for all sources).
[53] Representative sources are [30, 43, 82, 165, 199, 358, 452, 463, 485] (see Table 14 for all sources).
[54] Representative sources are [30, 51, 82, 236, 322, 348, 358, 463, 489] (see Table 15 for all sources).
[55] Representative sources are [155, 306, 341, 409] (see Table 16 for all sources).

For a system to be *fair*, the influence of protected attributes (e.g., gender or ethnicity) in the system's decision-making processes must be appropriately limited or controlled. Machine explainability can aid in this regard by providing means to track down factors that may have contributed to unfair decision-making processes, either to help eliminate such factors [54], to support mitigating them [304], or at least to raise awareness of them [399].

This complements Motivation 1.2 of machine ethics. Where machine ethics requires that values used in decision making are made explicit, machine explainability can help communicate those values to humans. Moreover, machine explainability can even help machine ethics here, by enabling the identification of the values in question.

Informed Consent[56] Informed consent is a legal concept in which regulators are interested and which is closely linked to autonomy. In essence, informed consent is about the consent that patients give to treatment in a healthcare setting. These patients should be sufficiently informed about a treatment so that they can accept or refuse it autonomously.

With the rise of ML systems supporting diagnosis and treatment recommendations in healthcare, informed consent is a desideratum that has extended from the everyday doctor–patient relationship to more complex situations involving artificial systems. Such systems promise to make more accurate predictions of an individual's health status, but with the drawback that these predictions are less traceable. The reduction in traceability threatens patients' ability to provide informed consent, as well as their autonomy.

Explainability can facilitate informed consent, similar to how it can facilitate *autonomy* (see above): it helps to equip users with the necessary knowledge for making informed decisions.

Legal Compliance[57] With the increasing proliferation of artificial systems, questions about their regulatory compliance emerged. Indeed, the proliferation has become so rapid that regulators are struggling to keep up with providing adequate laws for the use and deployment of these systems, and legal gaps have started to emerge. Moreover, the opaque nature of many systems makes it difficult to verify compliance with existing legislation.

Machine explainability can directly and indirectly contribute to legal compliance [189]. On the one hand, machine explainability can contribute indirectly, by enabling deployers and other stakeholders to verify whether a system satisfies certain lower-level desiderata, such as safety and non-discrimination, that are essential for legal compliance. Since deployers bear some responsibility for the systems they put into operation, they must ensure that these systems are at least legally sound to guarantee a lawful use.

On the other hand, the European General Data Protection Regulation and its often discussed *right to explanation* (arguably) explicitly require explanations (see [193] for a discussion). Other upcoming regulations for artificial systems (e.g., the upcoming European AI act) also

[56] Representative sources are [173, 342, 373] (see Table 17 for all sources).
[57] Representative sources are [1, 4, 72, 73, 173, 189, 193, 302, 442, 503] (see Table 18 for all sources).

contain formulations that seem to ask explicitly for explanations. Accordingly, explainability may also contribute directly to legal compliance.

Morality[58] We have already discussed the relationship between machine ethics and machine explainability in detail in the first part of this thesis, in Section 3.2. Here, we will only briefly present some considerations already made. In essence, for a system to be *moral*, its decision-making processes must be based exclusively on morally permissible considerations.

Autonomous Vehicle #11

According to certain moral theories, an autonomous car in a dilemma situation should never let affected parties' age contribute to its decision-making processes [313].

Explainability can contribute to achieving moral decision making [417] and, in particular, to so-called *ethical AI*. First, explaining a system's choice can help ensure that moral decisions are made [399]. Furthermore, providing explanations can itself be considered as morally required [292, 483]. Finally, as discussed earlier, explainability can contribute to *fair* decision making (see paragraph above), which is an essential aspect of *moral* decision making.

Performance[59] There are many ways in which a system can achieve better *performance*, and, just as with effectiveness and efficiency, we distinguish between a system side and a user side of performance. On the system side, *effectiveness*, *efficiency*, and (predictive) *accuracy* can be considered as performance measures. On the user side, "performance" can be understood as the quality of interaction between the user and the system. The better users can interact with a system, the better they, the system, and the user–system combination can perform.

Explainability can positively influence the performance of a system, among others, by helping developers to improve the system through an increased debuggability [393]. In this way, explainability positively affects other performance measures, such as *effectiveness*, *efficiency*, and *accuracy*. In particular, the case of *accuracy* is discussed in detail above.

For users, providing insights about a system, its functioning, and its outputs is a fruitful way to improve user–system interaction [54, 510]. Similarly, explanations can help improve user *performance* in problem solving and other tasks [274].

On the system side, increasing a system's overall performance can also increase its moral performance, which mitigates Risk 2.2.1 of machine ethics. The better a system's performance is and the easier it is to increase its performance, the more likely it is that its moral performance is also high or can be further optimized. Moreover, better performance on the user side contributes to Motivation 2.1 of machine ethics. With better overall user decision making based on a system's outputs, the system's users will likely also make better moral decisions.

[58] Representative sources are [4, 54, 72, 165, 175, 189, 236, 322, 373, 391, 399] (see Table 19 for all sources).
[59] Representative sources are [54, 123, 137, 173, 199, 231, 274, 366, 503, 510] (see Table 20 for all sources).

Persuasiveness[60] Basically, persuasiveness is about the likelihood that the outputs of an artificial systems will be leveraged by users to take further action. Persuasiveness is vital for deployers because they want the systems that they bring into use to make a difference for the users.

High persuasiveness can lead to good results if the system's recommendations are better than those made by humans. Too much persuasiveness, however, can also lead to bad results for users, for example, if a company uses systems to nudge persons according to its will [114].

Explainability can contribute to *persuasiveness*, since it can increase the acceptance of a system's decisions and the likelihood that users adopt its recommendations [358].

Privacy[61] Modern systems are notorious for collecting large amounts of data from their users, and for thereby violating their privacy. However, users have an interest in preserving their privacy so that their data cannot be exploited (whether by large companies or by individual scammers).

Explainability can have a positive impact on privacy, as disclosing information can help users figure out which features are correlated with sensitive information that should be confidential [233].

Reliability[62] Reliability is an essential desideratum of deployers, as they have an interest in the system which they bring into use working well and safely. Accordingly, this desideratum is also of interest to the users, since they want systems that perform as needed.

In the same way in which it can support *safety*, explainability can also help *reliability*, for instance, as it helps to fix bugs and contributes to *debuggability*.

Reliance[63] Similar to *confidence* and *trust*, it is an essential desideratum of deployers that users rely on the systems they work with. For more information on the differences between these three concepts, see the paragraph above on *confidence*. For more information on the relationship between reliance and explainability, see the paragraph below on *trust*.

Responsibility[64] The ability to determine who is blamable or culpable for a mistake is called *accountability* or *responsibility*. With the advent of artificial systems, a responsibility gap has emerged [323, 390]. For instance, when the use of an artificial system harms a person, it may not be clear who is responsible because there are many parties that may have contributed to the harm. Opacity of artificial systems only exacerbates this problem.

A person acting on the outputs of an artificial system may not (be able to) know that those outputs were erroneous, so blaming the person for the ensuing problems might not take the

[60] Representative sources are [51, 111, 137, 165, 180, 339, 348, 358, 463] (see Table 21 for all sources).
[61] Representative sources are [54, 189, 233, 236] (see Table 22 for all sources).
[62] Representative sources are [1, 189, 362, 427] (see Table 23 for all sources).
[63] Representative sources are [72, 93, 231, 388] (see Table 24 for all sources).
[64] Representative sources are [130, 339, 342, 358, 365, 373, 388, 397] (see Table 25 for all sources).

system's contribution into sufficient account. Overall, regulators aspire to avoid situations where existing legislation is difficult to apply or where no one is (or feels) responsible for a mistake.

Explainability can restore responsibility by making the errors and causes of unfavorable outcomes identifiable and attributable to the parties involved. Responsibility is an exemplary case of a desideratum associated with explainability, and we will explore it in Section 9.3.

Machine explainability, with its potential to support responsibility attribution, mitigates the risk, associated with machine ethics, of making the attribution of responsibility more difficult (Risk 2.2.2). Even if there are more candidates who might bear responsibility, for example, because new people devise moral codes for machines, machine explainability promises that it is possible to determine exactly which party should be held responsible for a failure.

Robustness[65] Robustness is desideratum of developers for ML-based systems. Basically, robustness is about the change in prediction across similar inputs. Ideally, this change is small; such a system is deemed robust. However, many ML-based systems suffer from poor robustness, and human-imperceptible changes in input can lead to large changes in prediction [192].

Explainability can contribute to robustness, for instance, through model optimization (for more information, see the paragraph above on *debuggability*).

Safety[66] Safety is a crucial desideratum across all stakeholder classes. Developers need to design safe systems so as to minimize the occurrence and severity of failures. Additionally, this limits the developers' accountability in case of failures since the appropriate safety measurements have been put in place. Users want only safe systems that do not expose them to danger. Deployers want safe systems to be put into use, as they hold the responsibility for many lives. Finally, regulators want safe systems to guarantee the best outcomes for society.

Many of the desiderata that we spoke about contribute to safety, and, thus, explainability contributes, too. Among others, controllability, debuggability, education, legal compliance, and robustness contribute to safety. Overall, explainability, thus, has a positive impact on *safety*, helping to meet safety standards [399], or helping to create safer systems [201].

Satisfaction[67] One desideratum that both developers and deployers have is that users are satisfied with a system. For the deployers, a system that does not satisfy its users is likely to be underused and, therefore, needs improvement. This is a reason for developers to make the system satisfying from the outset, so as to avoid having to redevelop it.

Satisfaction has many facets that explainability can contribute to. For example, explainability can have a positive effect on the usefulness of a system or recommendation to users

[65] Representative sources are [54, 73, 187, 322] (see Table 26) for all sources.
[66] Representative sources are [30, 153, 187, 236, 280, 391, 399] (see Table 27 for all sources).
[67] Representative sources are [51, 73, 137, 173, 199, 225, 348, 358, 388, 463, 478] (see Table 28 for all sources).

[502] (see also the paragraph below), which contributes to the perceived value of a system, increasing users' perception of the competence [382] and integrity [274] of a system, and leads to a more positive attitude towards the system [124].

Science[68] In the natural sciences (e.g., biology, chemistry), the use of systems based on ML has increased significantly in recent years. These systems are fed with large amounts of data (e.g., from experiments), to detect associations between variables and to make predictions about future experiments. Overall, many interesting systems could be created in this way (e.g., *AlphaFold* for protein structure prediction, see [256]).

However, the ultimate goal of science is not to prove association, but causality. Ideally, researchers strive to form hypotheses based on the discovered associations, and to test the hypotheses to gain insights into causality. Yet, because modern ML-based systems exhibit a high degree of opacity, this process is exacerbated; it is often almost impossible to form useful hypotheses.

Explainability can support scientific discovery [399]. For example, by making the decision patterns in a system comprehensible, knowledge about the corresponding patterns in the real world can be extracted. This can provide a valuable basis for forming useful hypotheses [304].

Security[69] Given the increasing dependence of humans on artificial systems, exploitation of these systems can lead to very negative consequences. Therefore, it is crucial to ensure the security of these systems.

Explainability is seen as a means of bridging the gap between perceived and actual *security* [373], helping users to understand the actual mechanisms in systems and adjust their behavior.

Higher system security helps to mitigate the potentially increased corruptability of systems due to machine ethics components (Risk 2.1.1). By enabling users to adapt their behavior according to a system's security, and by helping developers to make systems more secure, corruption is made more difficult, hence alleviating the potential consequences of corruption.

Transferability[70] Transferability is a desideratum of developers in the ML domain, and concerns the possibility of reusing a learned model in new contexts (in other words, it concerns the portability of ML models). Training modern ML models is a very resource-intensive process, taking quantities of, for instance, time and energy that should be reduced when possible. However, training is often specific to definitive application domains, making it difficult to transfer models to new contexts.

Explainability can help in this regard by enabling the identification of the context from, and to which, the model can be transferred [109]. Furthermore, explainability can facilitate the task of clarifying the boundaries that might affect a model, allowing for a better understanding and implementation in new contexts [54].

[68] Representative sources are [4, 54, 73, 179, 341, 399, 441, 442, 454, 503] (see Table 29 for all sources).
[69] Representative sources are [236, 322, 373, 503] (see Table 30 for all sources).
[70] Representative sources are [54, 234, 497] (see Table 31 for all sources).

Transparency[71] Transparency is a desideratum about which one could discuss whether it is facilitated at all by explainability. In the ML community, *transparency* is often used synonymously with *understandability*. If one considers the term as such, however, a system being transparent simply means that one has access to all of its parameters and processes.

In this latter sense, making a system explainable does, in principle, not change its transparency. In the former sense, however, explainability can contribute to higher system transparency [111], which is also illustrated by our model (see Figure 11). For more information, see the paragraph below on *understandability*.

Trust[72] The desideratum of adequately calibrating *trust* (*confidence*, *reliance*) in systems is one of the most discussed in the literature [264]. Both undertrust and overtrust have a negative impact on the appropriate use of systems [301]. In the case of undertrust, users may constantly attempt to monitor a system's behavior or even interfere with a system's processes, thereby undermining the effectiveness of human–system interaction [368]. In the case of overtrust, people might use a system without questioning its behavior [230, 367, 368], which can also undermine the effectiveness of human–system interaction, as people rely on the system's outputs even when they should question them [285, 301].

Trust is a complex desideratum, and the relationship between trust and explainability is often overestimated, just as is the case for acceptance (see Section 3.2.2). In general, explainability has the potential to allow users to appropriately calibrate their trust in artificial systems [232], and the same is true for the related desiderata of *confidence* and *reliance*.

However, there are also more complex cases. Suppose the desideratum is not to calibrate trust in a system, but to increase trust in it (as it often can be found in the literature, see [264]). Acquiring a higher degree of understanding will increase a stakeholder's epistemic satisfaction of this desideratum (i.e. they can better assess whether and to what extent to trust the system), but the substantial facet (i.e., their actual trust) may be negatively affected.

The reason for this is as follows. When a stakeholder still possesses a low degree of understanding, they are likely to be unaware of problematic features a system has in certain contexts (e.g., in complex environments) or for certain kinds of input data (e.g., noisy inputs). So, with a low degree of understanding, a stakeholder is likely to trust a system (although inadequately) [72, 274]. In contrast, a stakeholder with a higher level of understanding is able to recognize or even explain the conditions under which a system tends to fail. Therefore, they are more aware of the system's problematic features, which may, consequently, decrease their trust in it [113, 274]. It should be noted here, however, that this is an advantage, as the blunt elevation of trust, without any justification, is not morally desirable.

[71] Representative sources are [43, 72, 137, 177, 189, 322, 373, 399, 442, 485] (see Table 32 for all sources).
[72] Representative sources are [72, 188, 393] (see Table 33 for all sources).

Trustworthiness[73] The concept of *trustworthiness* is only vaguely defined [331]. For example, the high level expert group on artificial intelligence (HLEGAI) initiated by the European Commission does not provide a general definition of trustworthiness, but merely suggests that trustworthy systems have three characteristics: they are *lawful*, *ethical*, and *robust* [172]. Without elaborating much on trustworthiness, the HLEGAI emphasizes the significance of trustworthy artificial systems by stating that the trustworthiness of systems is imperative for the realization of potentially large social and economic benefits.

Regulators such as the EU, as well as previous research on AI that calls for trustworthy systems (e.g., as described in [175]), agree that explainability is a central way to test and facilitate the trustworthiness of systems [172, 175]. In particular, explainability can contribute to the three characteristics described by the HLEGAI, as we explore in the paragraphs on *legal compliance*, *morality*, and *robustness*. Moreover, philosophical notions of trustworthiness often have a justification component (i.e., a stakeholder must be justified that a system works well) [331]—and explanations are a good way to provide justifications [264].

Understandability[74] Sometimes understandability itself can also be a desideratum. As previously outlined, our model assumes that understanding is a step from explainability approaches to the satisfaction of other desiderata. Of course, this does not preclude understandability itself from being a desideratum. In this case, the relationship between explainability and desiderata satisfaction is very simple, and as outlined above: by receiving explanations about (certain aspects of) a system, a person's understanding of the system is facilitated.

Usability[75] In many cases, a system is more usable if it provides meaningful information about the reasons underlying its outputs. This information can help users to adequately link their knowledge and assessment of a given situation with the information used by a system. It can help them to make decisions more quickly, or it can increase decision quality [170].

Explainability can influence the *usability* of a system by providing required information. Explanations can increase a system's ease of use [358], leading to more efficient use [508], and making it easier for users to find what they want [464].

Usefulness[76] All of the factors mentioned in the previous paragraph can contribute to the usefulness (another important desideratum of users) of a system. Usefulness is important in high-stakes scenarios where a user makes a decision based on a system's recommendations. In these situations, it is imperative that the recommendations are actually useful.

[73] Representative sources are [72, 73, 172, 188, 201, 306, 339, 393] (see Table 34 for all sources).
[74] Representative sources are [54, 99, 339, 393, 399] (see Table 35 for all sources).
[75] Representative sources are [1, 51, 137, 199, 236, 358, 373, 388, 442, 511] (see Table 36 for all sources).
[76] Representative sources are [165, 180, 199, 358, 388, 442] (see Table 37 for all sources).

Verifiability[77] To check whether a system works as intended is commonly referred to as *verification*. Verifiability is a critical desideratum for developers, since they must ensure that the systems they produce do, indeed, work as intended. Program verification can range from empirical testing to formal proofs, both of which approaches have received intense interest in the research community. However, new challenges are introduced by traditional verification techniques not being applicable to ML-based systems; different paradigms are required.

Medical-Care Robot #29

Caruana et al. describe an example of potential issues with verification arising from opaque systems [98]. They discuss a system that predicted asthmatic patients to have a lower risk of dying from pneumonia than non-asthmatic patients. This prediction arose because the training data were systematically biased: Asthmatic patients were better monitored than other patients and, consequently, they died less frequently from pneumonia, even though they were a high-risk group. If the system in question (e.g., as part of our medical-care robot) were used to assess the treatment urgency of patients, this could have fatal consequences—the system would assign low treatment urgency to the high-risk group. For more examples of this kind, see [296].

Better insight into the system's decision-making process through the use of certain explainability approaches can help developers to identify, and potentially correct, such mistakes [344]. Not only can explanations help developers ensure the correctness of a system's knowledge base [137], they can also help users to evaluate the accuracy of a prediction [510]. Overall, when people have access to the processes and data that lead to an output, they can better judge whether the output is appropriate. Explainability is key to this.

By supporting verification, machine explainability provides a third factor that can help mitigate the potentially increased corruptability of systems (Risk 2.1.1). In essence, by enabling people to verify that the systems do what they should do, it is possible to exclude that they are being used maliciously.

The various desiderata arising from the five classes of stakeholders are manifold and diverse; presenting a comprehensive list lies beyond the scope of the current thesis. The above examples have been drawn from numerous sources, all of which postulate explainability as an an important ingredient for satisfying these desiderata.

However, in previous research, it has mostly remained unclear how explainability is supposed to satisfy each of the desiderata coming from the various stakeholder classes. With our conceptual model, we offer clarification in this regard.

[77] Representative sources are [66, 341, 344, 409] (see Table 39 for all sources).

9.2.2. Reasons against Machine Explainability

Having discussed the overwhelming number of desiderata that can be satisfied in various ways by explainability (thereby emphasizing its need), we now turn briefly to the desiderata that might be frustrated by it (thereby hinting at its drawbacks).

Basically, some of the desiderata that may benefit from explainability may also suffer from it. For some of them, we have already briefly touched upon this. Performance and, relatedly, accuracy, effectiveness, and efficiency offer paradigmatic cases (see [126] for a discussion). These desiderata may suffer on the system side because making a system explainable may require additional computational resources which lower performance. On the user side, poorly designed explanations can lead to confusion, which lowers user performance [96].

Other desiderata that could suffer are security and privacy. If different stakeholders are provided with more information about the system and how it works, they could maliciously exploit this information. However, security through obscurity is considered a poor design decision among experts [335], which is why open software is often preferred.

Considering that the number of desiderata that might benefit from explainability is far greater than the number that might suffer from it, and considering that even this suffering is not inevitable, this reason against machine explainability alone seems rather weak. Therefore, we will cite some other reasons that might speak against making systems explainable.

One reason sometimes cited is that machine explainability, as a research discipline, is too inconsistent to be fruitfully pursued [397]. We have addressed some of the problems associated with this reason above (Section 8.3.2; e.g., problems with the use of terms such as “explainability” and “interpretability”). A more detailed discussion regarding such inconsistencies and possible solutions to them can be found in [447].

Another reason portrays research on explainability as misdirected [280, 365]. According to this line of argumentation, the time spent on researching machine explainability would be better spent on other endeavors, such as finding ways to directly satisfy individual desiderata. One claim is that there is no guarantee that explainability will ultimately contribute sufficiently to satisfying these desiderata, so the whole endeavor may fail. Given the immense resources invested in research on explainability, this could prove disastrous.

However, this reason can also be easily alleviated. Research on explainability has already produced impressive result. For instance, when it comes to fairness, many biased ML-based systems could be identified (see, e.g., [272, 297]). Furthermore, for some desiderata, explainability seems the best and most direct way to satisfy them (e.g., responsibility; see below). Accordingly, pursuing explainability research is worthwhile because it does satisfy the desiderata and there is no alternative method that seems nearly as promising.

9.3. The Responsibility Desideratum

In the above descriptions of the desiderata and the purported influence of explainability on them, we have remained rather general. On the one hand, detailed analyses of so many desiderata would simply deviate too far from the crux of this thesis. On the other hand, for many desiderata, the detailed connection to explainability remains to be explored. In this section, we will outline one such connection, namely, to responsibility.

Even with such a focus on a single desideratum, we must limit ourselves to one specific example, since the attribution of responsibility is such a complex issue. More specifically, we will focus on a case where a human makes a decision based on results from a system. This is a typical so-called *human-in-the-loop* case.

9.3.1. The Epistemic Condition for Moral Responsibility

Prima facie, a human in the loop is an excellent candidate to bear responsibility. However, there are some requirements that have to be fulfilled to properly allocate responsibility to them. This is where the demand for explainability comes into play. As Floridi et al. put it, ensuring "that the technology—or, more accurately, the people and organizations developing and deploying it—are held accountable in the event of a negative outcome [...] would require [...] some understanding of why this outcome arose" [175, p. 700]. To gain such an understanding, the human in the loop must have access to an explanation of an artificial system's recommendation, and possibly even its entire operation, at the time of the decision.

The basis for this idea lies in a necessary condition for moral responsibility that is widely recognized in philosophical debate—the *epistemic* condition (EC) [354, 405]. According to this condition, an agent is directly morally responsible for an action only if they are aware, or in a position to be aware, of (a) what they are doing, (b) what the (probable) consequences of their action are, (c) what moral significance it has, or (d) what alternatives are available to them.

(EC) An agent is directly morally responsible for their action or decision only if they have sufficient epistemic access to it. That they have sufficient epistemic access to it entails at least that they are in a position to know the action under relevant descriptions.

Example #21

For example, an agent who flips the switch to turn on the light, and who thereby electrocutes their neighbor by an unfortunate and unforeseeable combination of circumstances is not directly responsible for the harm caused.

To make this condition more tangible, we resort to a coarse-grained view of actions, according to which the description of an action can be singled out among a number of

different ones [32, 141] (note that this stands somewhat in tension with the assumptions in our framework, where we have reduced similar actions to paradigmatic ones).

Example #22

In the example, the agent's action can be described as flipping the switch, as turning on the light, or as electrocuting the neighbor. Since the agent is not in a position to be aware that their action is to electrocute their neighbor, they are not directly morally responsible for it under that description, though they may still be responsible for flipping the switch.

The epistemic condition on moral responsibility can be used to provide two motivations for artificial systems to be explainable. Let us make these explicit by means of examples.

9.3.2. Explainability for Bridging the Responsibility Gap

In particular, the first motivation is introduced using an initial case (Hiring System #2), and the second is introduced using an extended version of that case (Hiring System #3).

Hiring System #2

Herbert, the human resources (HR) manager, is a human in the loop and makes the final hiring decision, but has no explanation for the hiring system's recommendation. Let us assume that, prior to using the system as support, Herbert was an HR manager who competently and responsibly made hiring decisions for his company, and that he will continue to do so, using an artificial system's output as one source of support.

Imagine that Herbert decides to exclude the application of April (a Black woman) because the hiring system recommended doing so. Imagine further that the system's recommendation is due to its bias against Black female applicants, but, since it is an accredited system, Herbert justifiably believes that it has no such problems.

In this example, Herbert is not responsible for discriminating against April—he is not to blame for being unaware of the system's bias. If he is to be responsible at all, he must be directly responsible, which requires that he is in a position to know what it is that he is doing, what its probable consequences are, and what its moral significance is.

If, as described, he does not have access to what moved the artificial system to make its recommendation, then his AI-supported decision will be made without his being in a position to know these things. Herbert is aware that he is rejecting April's application, and so he is aware of his action under that description. However, he is not in a position to know that what he is doing, under another description, is discriminating against her. He is also not in a position to know that he unfairly rejects her application and that this is an act of moral wrongdoing. Consequently, he is not morally responsible for discriminating against April.

Once a meaningful explanation of the recommendation is available to the decision maker, we can more easily bridge the responsibility gap. Suppose that the system discriminates against April on the basis of her race and gender. If Herbert has access to this fact, then he also has access to the fact (in other words: he is in a position to know) that rejecting her application on this basis is to discriminate against her; and that it is unfair and an act of moral wrongdoing.

Even in the case where the system discriminates against April based on a learned correlation involving some otherwise innocuous proxy variables such as April's alma mater, her hobbies, and her zip code, explanations may allow Herbert to gain the right kind of epistemic access. This is because the proxies will typically be either suspicious or seemingly irrelevant. In either case, Herbert should doubt the system's recommendation: If the system indicates that it considers the combination of April's alma mater, hobbies, and current zip code to be particularly crucial, this may catch Herbert's attention. He might wonder: is this not one of the historically Black colleges and universities? And is that not a primarily Black neighborhood?

In any case, an explanation allows Herbert to become suspicious and to pay particular attention to the role played by other factors. Herbert can then check, if necessary, whether candidates with otherwise similar profiles are rated similarly. In this case of proxy-based discrimination, Herbert may not be sure that discrimination is present, but with sufficient background knowledge and awareness of the danger of discrimination by models, he can develop an initial distrust and at least begin to consider that other descriptions of the situation might be relevant. He is, therefore, in a position to know at least that a decision following the system's recommendation may very well be discriminatory. Thus, even if explanations do not guarantee that the EC will be fulfilled in all cases, they clearly facilitate its fulfillment.

9.3.3. Explainability for Resolving Cases of Disagreement

Let us turn to our second motivation. At least in one way of further spelling out Herbert's situation, his epistemological situation is even worse than has become clear so far.

Hiring System #3

Imagine that, at the end of a lengthy selection process, Herbert is presented with a list of applicants that the artificial system ranks as the top three; the system recommends keeping them in the running for the position. April did not make the list, but is in the top ten. However, Herbert, who went through the top ten applications independently, placed her among the top three applicants beforehand. So we have a case of disagreement between the system's recommendation and Herbert's initial judgment. Since there is no explanation for the system's recommendation available, Herbert cannot reasonably resolve the disagreement.

This scenario shows that, when a decision maker cannot tell why an artificial system provided the recommendation it did, there can be situations, especially situations of disagreement between system and decision maker, in which the decision maker cannot tell whether the system's recommendations bring them closer to their goals. As a consequence, they are unable to guide their decisions so as to pursue these goals, or to execute their intentions in acting. This gives rise to a particularly threatening way in which an agent lacks epistemic access to their action, and, thereby, also lacks moral responsibility for it.

Hiring System #4

Suppose that Herbert's own assessment of April's qualities is due to good, but not conclusive reasons—she has more relevant work experience than most, received great grades in her studies at Yale, speaks a foreign language which is useful but not absolutely necessary for the job, and has work experience abroad. (By saying that his reasons are not conclusive, we mean that they are sufficiently weak that he can reasonably question his own judgment if the system gives a contrary recommendation.)

On the other hand, the system was accredited as reliable by a trustworthy watchdog organization, though Herbert is aware that systems of this kind may have hidden bugs or biases. In this situation, the system's countervailing recommendation leaves open both the possibility that Herbert correctly assesses the situation and the system is mistaken, and also the possibility that the system has a superior understanding of the situation and Herbert has it wrong.

In the first possibility, the system's recommendation could be due to some kind of bug, or to its bias against women of color; in the second possibility, the system's recommendation could be due to having information that Herbert does not have, or that it detects patterns that Herbert overlooked.

Suppose that the system relies on all of the reasons that led Herbert to see April among the top three applicants (her excellent grades at Yale, her foreign language skills, etc.). However, it has found that applicants with these qualifications, overall, tend to move on very quickly to other, better jobs. So the system detects a pattern that turns what would otherwise be good reasons for hiring an applicant into a reason against hiring them.

This example illustrates that, in a particular situation, Herbert may be unable to tell whether he is in one of two relevant cases:

Case 1 The system's recommendation is mistaken and Herbert's assessment is right.

Case 2 The system's recommendation is correct and Herbert's assessment is wrong.

Given that the two cases are indistinguishable to Herbert, he cannot reasonably resolve the disagreement. This is because he cannot compare or reconcile his own and the system's

reasons for or against keeping April in the running, and, thus, cannot figure out which reasons are superior, for instance, by weighing them against each other. Consequently, if he decides to keep April on the shortlist, this decision is arbitrary; but if he decides to exclude her from the shortlist, that decision is also arbitrary. The lack of access to the system's reasons undermines Herbert's ability to arrive at a well-founded, all-things-considered judgment about which applicants to keep in the running.

In light of this inability, Herbert is, then, unable to competently pursue his goal. Suppose he is genuinely trying to find the best candidate for this prestigious, responsible position in his company. Since he is unable to discern the proper means for doing so—keeping April in the running or excluding her—he is unable to respond to pertinent reasons in pursuit of his goal. In other words, he cannot properly guide his decisions in light of his goals, so as to execute his intentions. This undermines his ability to find the best candidate or to achieve various related goals.

Imagine that Herbert is instead trying to harm the company by hiring an unsuitable candidate. Again, since he cannot tell whether it is his or the system's assessment of April that is right, he is unable to tell whether excluding April would be a good means to achieving this goal, which, again, undermines his ability to guide his hiring decision in response to pertinent reasons.

In both of these scenarios, Herbert is epistemically impaired: He cannot know any of his options under the relevant descriptions. He cannot tell whether he wrongs April by following the system's recommendation, but he also cannot tell whether his decision, if he sticks to his own initial assessment, can be described as harmful to his company. In this version of the scenario, then, Herbert's access to his decision is undermined more severely. Because of this wider-ranging epistemic disconnect, Herbert is not directly morally responsible for his AI-supported decision.

Of course, one might object that cases of disagreement are insignificant outliers. Typically, the decision maker will agree with the system's recommendation. However, this objection renders the use of artificial systems as decision support obsolete. If the system's recommendation allows for well-founded decision making only where it supports what the decision maker would choose anyway, then there is no point in combining an artificial system with a human in the loop for the hiring decision. From Herbert's perspective, adding the system does not improve his decision making; from the perspective of the company, keeping a human in the loop provides no advantage over employing a fully automated system.

The system can lead to better decision making precisely when the decision maker disagrees, and there is room for changing his mind. So, it is precisely when it matters—when the decision maker's reasons are not conclusive, and the system's recommendation is potentially better—that the system undermines the decision maker's epistemic access to their decision, and, thus, their moral responsibility.

9.3.4. The Dilemma of Lacking Explainability

Without explainability, we face a dilemma for human-in-the-loop scenarios: Either it is pointless to have the system provide a recommendation to the human decision maker (in cases where human and system agree, or when the decision maker has conclusive reasons anyway), or the lack of explainability undermines their epistemic access to their decision and, thus, the moral responsibility that the human in the loop is supposed to bear (in cases where human and system disagree, while the human has non-conclusive reasons).

Now the second horn of this dilemma is because the decision maker has no access to why the artificial system provided a particular recommendation. If they had a suitable explanation for the system's recommendation at their disposal, so that they could compare their reasons with the system's reasons, they would be in a better position to find out whether it is the system's or their own assessment of the situation that is correct. So, they would be able to resolve the disagreement in a non-arbitrary way and, thus, be able to make the hiring decision that best suits their goal (to find the right—or wrong—person for the job). Overall, they would be in a position to know their decisions and actions under the relevant descriptions.

We conclude that, in many cases of disagreement, where the decision maker's reasons are non-conclusive, they are in a position to bear direct responsibility for their decision just in case they have a suitable explanation of the system's recommendation available.

Generally, it can be stated that a human decision maker requires explanations. These enable responsible AI-supported decision making by enabling an agent to meet the EC. On the one hand, this allows for bridging the responsibility gap that has arisen with the introduction of artificial systems into human decision-making processes. On the other hand, it allows for the resolutions of disagreement between machines and humans.

Through these illustrations, one can see how explainability can be specifically connected to a desideratum. Furthermore, one can see that the effort to establish such a connection is very high. Yet, we have looked at only a very limited example. To explore the whole range of responsibility in the context of explainability would require a lot more space and additional research. Nonetheless, the example discussed nicely illustrates the benefits of explainability, and motivates further research on the connection of explainability to desiderata.

10. What Makes for a Good Explainability Approach?

At this point, we have a good understanding of the various desiderata associated with machine explainability. Now, recall our model (Figure 11). Desiderata satisfaction is intended to come through increased stakeholder understanding. To this end, researchers in the field of machine explainability are devising various approaches to provide explanatory information. So far, we have not commented much on these approaches. In this section, that will change.

We will first review the main distinctions between these approaches, and present some exemplary ones. In order to assess the quality of the approaches, we will then devise evaluation criteria for explanatory information, and apply them to the information produced by the presented approaches.

10.1. Exemplary Explainability Approaches

In order to get a better idea of explainability approaches, we will discuss some exemplary ones. These examples represent only a selection of well-known approaches and should in no way be taken as representative of the field as a whole. The number of proposed explainability approaches is huge, and it is difficult to keep track of them all [54, 201, 447]. That being said, in order better contextualize the presented approaches, we will first introduce the main distinctions regarding them (see [442] and [447] for more distinctions).

10.1.1. Important Distinctions Concerning Explainability Approaches

Explainability approaches can take many guises, and the literature commonly distinguishes them into two families (see, e.g., [4, 37, 54, 201, 442]): *ante-hoc* and *post-hoc* approaches.

Ante-hoc approaches aim at designing systems that are inherently transparent and explainable. They rely on systems being constructed on the basis of models that do not require additional procedures to extract meaningful information about their inner workings or their outputs. For example, decision trees, rule-based models, and linear approximations are commonly seen as inherently explainable (given they are not too large) [201, 341]. A human can, in principle, directly extract information from these models in order to enhance their understanding of how the system works, or of how the system arrived at a particular output.

Post-hoc approaches aim not at the design process of a particular system, but at procedures and methods for extracting explanatory information from a system's underlying model which need not be inherently transparent or explainable in the first place [201, 306, 442]. Post-hoc approaches are, for example, based on input–output analyses, or on approximating opaque models by models that are ante-hoc explainable.

One can distinguish post-hoc approaches that work regardless of the underlying model type (so-called *model-agnostic* approaches) from ones that only work for specific (types of) models (so-called *model-specific* approaches). Model-agnostic approaches aim to deliver

explanatory information about a system solely by observing input–output pairs [201, 393, 442]. Model-specific approaches do so while also factoring in specific features of the model at hand (e.g., by creating prototype vectors in a support vector machine) [201, 442].

Finally, previous research distinguished the *scope* of an explainability approach. Some approaches provide information about only single predictions of the model [201, 393, 442]. The scope of these approaches is *local*. Local approaches often offer visualized prototype outcome examples (see, e.g., [270, 271]). The more general type of approaches has a *global* scope. Global approaches are designed to uncover the overall decision processes in the model [201, 442]. Here, the usual way to provide this information is by approximating complex models with simpler ones that are inherently explainable.

10.1.2. Perturbation-Based Approaches

One big family of explainability approaches is the perturbation-based ones [447]. The idea behind perturbation-based approaches is to change the inputs of a system to observe the change in outputs. The information obtained from this observation can either be used directly or processed further. We will discuss three approaches that belong to this category below.

Change impact analysis (CIA) One very straightforward perturbation-based approach is CIA. This approach originally stems from software development, and is most commonly described as "identifying the potential consequences of a change, or estimating what needs to be modified to accomplish a change" [75, p. 3]. In the case of artificial systems, "change" must be understood primarily as the change in a parameter (or several parameters) used as an algorithm's input, which, in turn, is used in the system to come to a result. The difference in output when changing an input constitutes the explanatory information in CIA.

Example #23

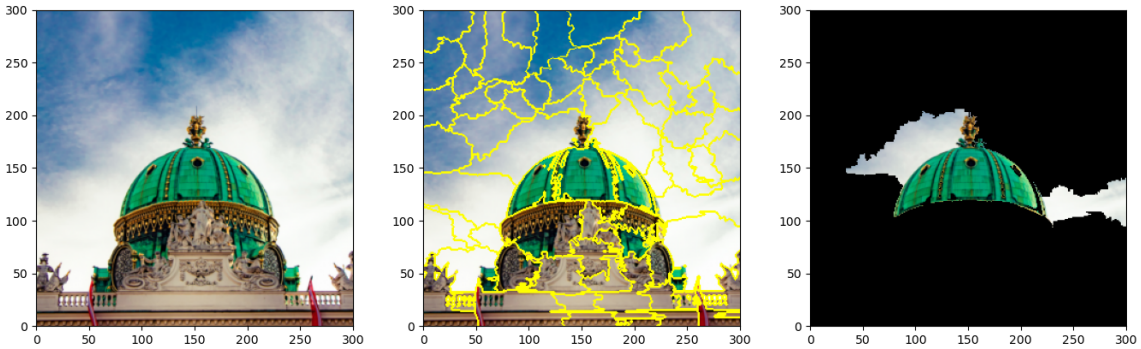
Avati et al. [45] suggest an ANN for estimating mortality rates (see also Example #25). To "explain" the network's predictions, its inventors suggest tweaking every parameter slightly to identify whether the prediction would change significantly. This "tweaking" is precisely the type of change upon which the CIA is based.

Robustness A similar approach involves looking for *robustness* guarantees. Particularly in ML-based systems, the behavior of trained classifiers can be hard to predict at times. Minor, human-imperceptible changes of an input can lead to significant changes in the output (in [192], Goodfellow et al. give a good example of what this can look like in ANNs). This behavior can be disastrous in systems that interact with humans.

Autonomous Vehicle #12

With observations like those made by Goodfellow et al. [192], it is plausible to assume that a slight noise in a sensor (for instance, a pixel error in or dirt on a camera) can lead to disastrous results in traffic sign recognition. Thus, the 50 km/h sign in a town could be mistaken for a 90 km/h sign, leading to an accident.

Guarantees concerning the robustness of a classification are, thus, desired. Research investigating this was made, for example, by Hein et al. [216]. Given a specific prediction from a classifier, they demonstrate a means of estimating the required change in an input that would have led to a change in the output. Overall, one could, thus, also count this as a type of CIA.



(a) Original image. (b) Interpretable clusters. (c) Decisive clusters.

Figure 13: Local interpretable model-agnostic explanation (LIME) visualized.

LIME A well-known explainability approach is LIME (proposed by Ribeiro et al. in [393]; see Figure 13). With its conception in 2016, LIME was one of the triggering factors for the renewed interest in machine explainability and has greatly influenced the field. The idea behind LIME (discussed below) is quite simple, which most likely played a large part in its popularity. As its name suggests, LIME is a local, post-hoc approach that is model-agnostic.

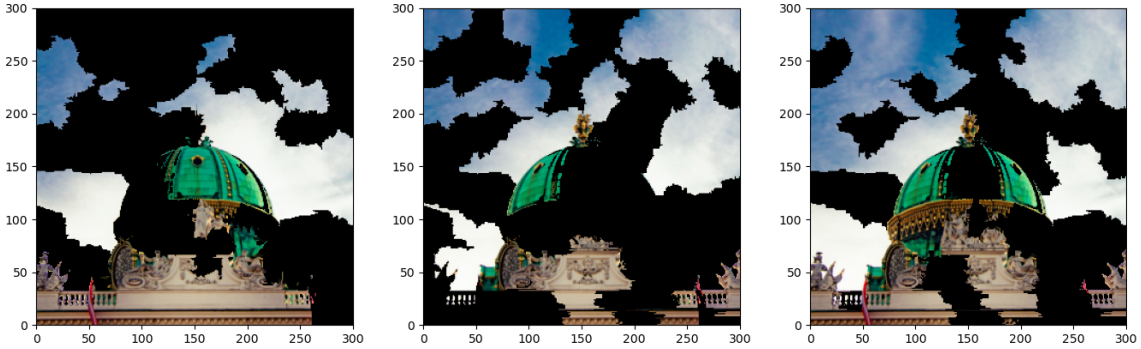


Figure 14: Different superpixel masks for LIME.

Let us describe LIME in simple terms. Since it is a local explainability approach, LIME aims to explain single predictions. Given such a prediction of a particular input (be it an image, a text, or something different; see Figure 13a for an example), the first step is to

segment the input into interpretable clusters (see Figure 13b). Subsequently, some of these clusters are replaced with dummy values (e.g., black in an image, see Figure 14). This is repeated many times to create a large number of new inputs that differ in the clusters that have been replaced. For all of these inputs, a prediction in the original model is made. Using the difference between these predictions and the original prediction, it is now possible to calculate which clusters contributed the most to the original prediction (see Figure 13c).

To understand on a more general level what LIME does, it is important to know that, in principle, all that an ML model does is compute a function. Such functions can become very complex, having a huge number of variables (e.g., 786432 for a model that takes images of 512×512 pixels with three color channels as input). LIME approximates this complex function for a given value by another, simpler function (e.g., a linear one). Figure 15 visualizes this: Assuming that the blue graph represents the function computed by an ML model, the red and the green graphs are each local, linear approximations to this function.

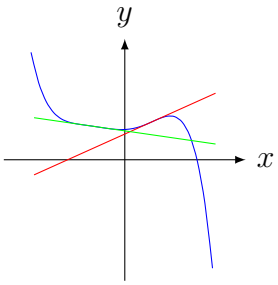


Figure 15: Linear approximations to a function.

With this information in mind, we can describe LIME in more complex terms. Ribeiro et al. use the following mathematical equation to express this explainability approach [393]:

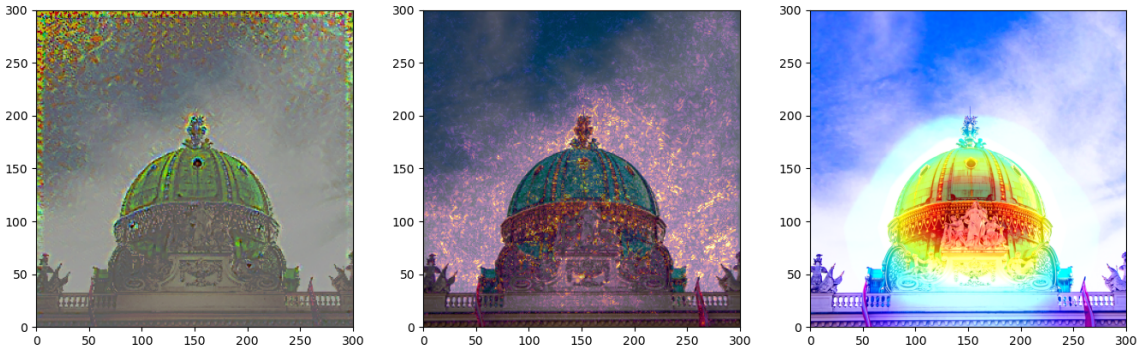
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g) \tag{1}$$

In this equation, g is the surrogate model (e.g., a linear regression model) out of the class G of potentially ante-hoc explainable models. Notably, Ribeiro et al. consider this model to be the explanation, readily presentable to an addressee with visual or textual artifacts. Moreover, Ribeiro et al. are aware of the fact that purportedly ante-hoc explainable models may not retain their intelligibility if they become too large. To prevent this from happening for LIME, they introduce $\Omega(g)$ as the measure of complexity of a model (e.g., the depth of the tree for decision trees, or the number of non-zero weights for linear models).

Other components are the original model f (e.g., a DNN) and the proximity measure π_x. As mentioned earlier, f is a function $\mathbb{R}^d \rightarrow \mathbb{R}$ (where d is the input's dimensionality), and $f(x)$ expresses the probability that an instance $x \in \mathbb{R}$ belongs to a particular class. The proximity measure defines locality around the instance x by being a way to calculate the distance between x and other instances. The final component is L, which measures the infidelity of g in approximating f in the locality defined by π_x. Overall, the core idea is, thus, to produce another ML model that is human-intelligible and locally fidelitous to the original model.

10.1.3. Saliency Maps

While LIME can serve as a prime example of a model-agnostic explainability approach, we will now briefly examine a number of mutually related, model-specific approaches. These approaches are related in that they all produce *saliency maps* while leveraging a convolutional neural network (CNN)'s structure. Moreover, some of them are even enhanced versions of others.



(a) Guided backpropagation. (b) Integrated gradients. (c) Grad-CAM.

Figure 16: Different approaches to generate saliency maps.

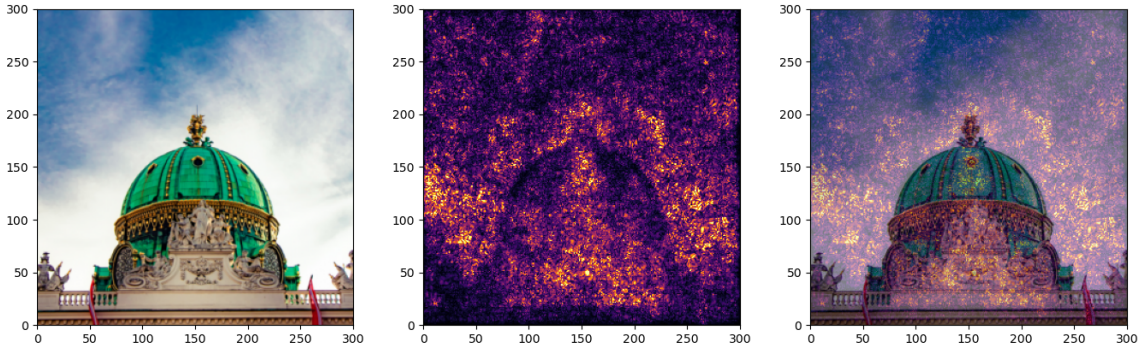
Before we come to the individual approaches, let us first give some background on saliency maps to better understand these approaches. Given a prediction for a certain input, a saliency map visualizes that input, highlighting which parts or features are most relevant to that prediction (see Figure 16). For simplicity, we will focus on images as inputs, although saliency maps can be created for other input formats as well. Additionally, while there are traditional computer vision algorithms for detecting saliency, we will deal exclusively with explainability approaches producing saliency maps in the context of CNNs.

To understand how these explainability approaches create saliency maps, we need to dive a bit into the specifics of how ANNs are created. In particular, we need to elaborate on the *backpropagation algorithm*. Basically, the backpropagation algorithm is what allows ANNs to be trained. For each training example, this algorithm determines how to adjust the individual weights between the neurons in an ANN to reduce the error (i.e., the difference between calculated and expected value), to increase the ANN's predictive performance. Roughly, this is done by computing the gradients of a function that is created based on said error. Moreover, this process is performed starting from the last layer of the network and finishing with the first, hence its name, "*back*propagation".

Based on the backpropagation calculations, the individual weights between the neurons are adjusted following each training example. In the best case, the ANN's performance reaches an adequate standard by being confronted with many examples and having the weights changed accordingly. At this point, the ANN has become useful for solving the problem posed to it (e.g., distinguishing pictures of cats from pictures of dogs).

Explainability approaches that produce saliency maps sometimes have "backpropagation" or "gradient" in their name because they visualize some form of gradient calculated by backpropagation. In general, gradients in ANNs indicate which inputs need to be changed the least to affect the prediction the most (for saliency maps, these inputs are pixel values). In this respect, these approaches are similar to LIME, whose goal is also to indicate what needs be changed the least to affect the prediction the most.

Figure 16a and Figure 16b may serve as examples of saliency maps: the most salient pixels are at the outline of the dome. This is to be expected, since having a semi-circular shape is essential to being a dome. Accordingly, the underlying model seems to work reliably for predictions of this type. How it behaves beyond that, however, this saliency map can not tell.



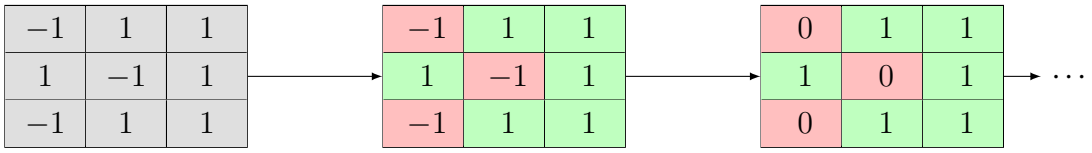
(a) Original image. (b) Activation mask. (c) Superimposed mask.

Figure 17: Vanilla backpropagation visualized.

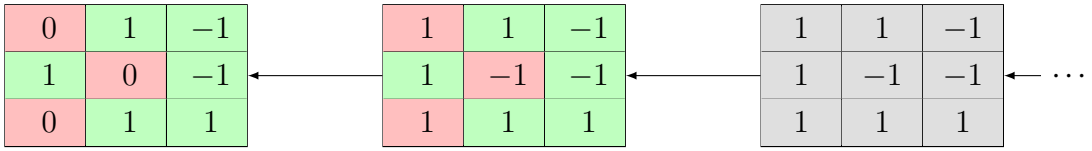
Vanilla Backpropagation The first work on saliency maps in CNNs was published in 2013 (viz., [430]) and proposed an explainability approach that is now often referred to as *"vanilla backpropagation"* (see Figure 17). Basically, this approach exploits the gradients used in the backpropagation error function. These gradients indicate whether a pixel in the image was important for a prediction or not. Accordingly, important pixels get highlighted.

To better put vanilla backpropagation into context, we need to comment on how it handles the so-called *ReLU function*. The ReLU function is an activation function in ANNs. Activation functions in ANNs try to simulate something similar to the excitation of human neurons. ReLU does this by disregarding negative incoming values, and scaling positive ones to one (see Figure 18a). In vanilla backpropagation, the values set to zero by the ReLU function are remembered and, during the backpropagation, also set to zero (see Figure 18b).

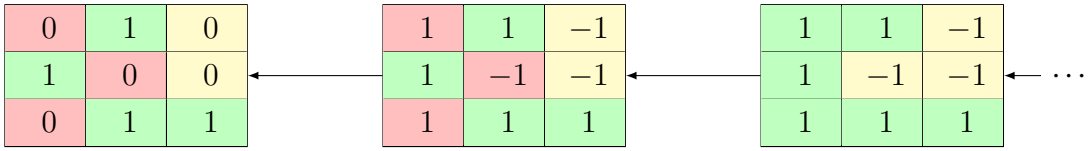
Unfortunately, vanilla backpropagation usually generates images that are extremely difficult to interpret, even by experts (see, e.g, Figure 17b). For this reason, there have been several proposals to improve vanilla backpropagation or to generate saliency maps in an altogether different way. In what follows, we will briefly discuss some of these approaches. However, as laying out the technical details for each of these approaches would lead too far from this thesis, we will only give a superficial account of how they work.



(a) The *ReLU* activation function.



(b) Vanilla backpropagation



(c) Guided backpropagation

Figure 18: Vanilla backpropagation and guided backpropagation explained.
When using an ANN to make predictions, the ReLU activation function disregards negative inputs (a). This knowledge is exploited by vanilla backpropagation by setting values that were negative to zero during backpropagation (b). Guided backpropagation goes further and additionally ignores negative gradients (c).

Guided Backpropagation An approach similar to vanilla backpropagation is *guided backpropagation* [450] (see Figure 16a), in which the focus lies on positive gradients and negative ones are ignored (see Figure 18c). The reason for this change of focus is that negative gradients indicate a decrease in probability of being predicted as a particular class (e.g., being a dome). Thus, these gradients tell us that certain pixels did not contribute to the prediction in question. However, as we are interested in what contributes to the prediction, we can disregard them.

Gradient-based techniques such as vanilla backpropagation and guided backpropagation suffer from various problems. Among others, there is the problem of so-called "vanishing gradients", which concerns very small gradients. Although such gradients might indicate important pixels, they tend to become smaller during backpropagation, ultimately vanishing.

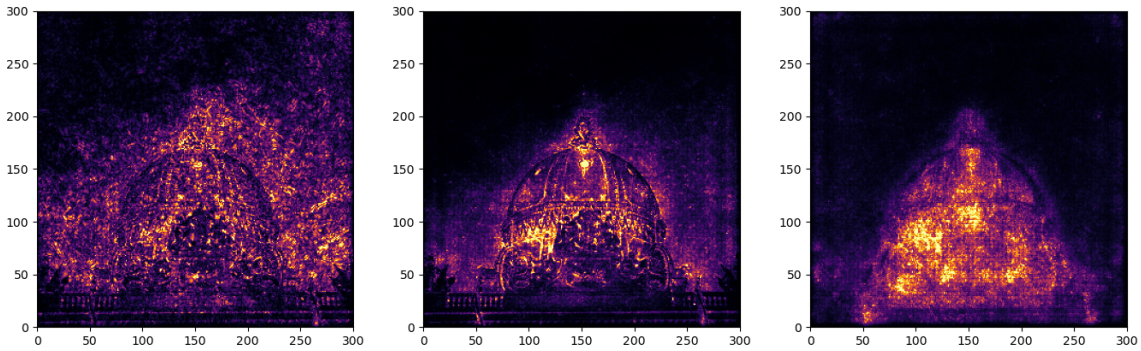
Integrated gradients A technique that attempts to circumvent this problem by computing the gradient differently is *integrated gradients* [455] (see Figure 16b and Figure 19a). It works by taking derivatives of the value for the predicted class with respect to the input features. This is done along a straight line from a given baseline (a representation of an input that reflects the absence of a signal, such as a black image) to the actual input.

All previous saliency techniques suffer from artifacts that can make the resulting masks very enigmatic. So, the question is how to reduce these artifacts.

SmoothGrad One way to reduce artifacts is to combine multiple outputs of the same approach (e.g., of vanilla backpropagation), such as with *SmoothGrad* [438]. The idea behind

SmoothGrad is as simple as it sounds odd: reducing noise by adding noise. What seems like a contradiction at first glance makes a lot of sense upon closer inspection.

SmoothGrad interpolates between multiple saliency masks, each of which is created by using one of the above approaches. This is done by applying the approach to slightly modified version of the original image. These images are modified by the addition of Gaussian noise. One type of noise (the artifacts) is, thus, reduced by adding another type of noise (the Gaussian noise in the modified images). As can be seen in Figure 19, the diffuse attribution masks of Vanilla Backpropagation (see Figure 17b) and Integrated Gradients (see Figure 19a) become very clear with SmoothGrad (see Figure 19b and Figure 19c, respectively).



(a) Original mask of integrated gradients.

(b) SmoothGrad mask of integrated gradients.

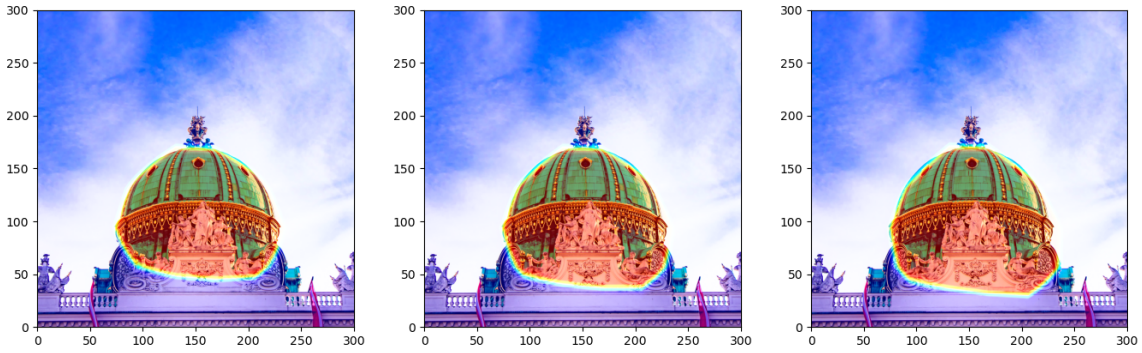
(c) SmoothGrad mask of vanilla backpropagation.

Figure 19: SmoothGrad for integrated gradients and vanilla backpropagation.

After these approaches, which rely primarily on the gradient, let us turn to some alternatives. To this end, we need to elaborate a bit on what distinguishes CNNs from other types of ANNs. CNNs make use of the name-giving *convolution*. Basically, convolution is a mathematical operation that allows the merging of two sets of information. In the case of CNNs, certain layers use convolution to create feature maps of input images.

Class Activation Maps CAMs A technique that uses the last of these convolutional layers to detect the features that contributed the most to a prediction is to create *CAMs* (proposed in [507]). Unfortunately, CAMs have specific requirements on the structure of the CNN to be explained. Thus, generalizations of CAMs have been developed.

Grad-CAM One such generalization is *Grad-CAM* (proposed in [420, 421]). The basic idea is to combine gradient-based insights with CAMs. In this way, the most important CAMs for a prediction can be detected, just as other saliency maps detect the most important pixels for a prediction. Typically, the results of CAMs and Grad-CAM are heatmaps that can be overlaid on top of the original image (see Figure 16c and Figure 20).



(a) Grad-CAM (b) Grad-CAM++ (c) Score-CAM

Figure 20: Approaches that are based on CAMs.

Other CAM-Based Approaches As one might expect after seeing this variety of approaches, Grad-CAM is also not without problems, so alternatives have been developed.

Grad-CAM++ (proposed in [104]) builds directly on Grad-CAM, slightly changing the way the derivatives are used to compute the most important CAMs (see Figure 20b).

Score-CAM (proposed in [486]) takes a different approach, combining the occlusion idea of LIME with CAM to compute importance scores for features (see Figure 20c).

Figure 20 visualizes the three CAM-based approaches we have discussed with a condensed heatmap. Although the difference is minimal, there is a recognizable shift of the highlighted zone from the upper left to the lower right.

10.1.4. Further Approaches for (Convolutional) Neural Networks

The examples for explainability approaches we have given so far use *feature-importance* attributions to explain black-box models (mostly CNNs). That is, the approaches highlight specific features, such as the (clusters of) pixels that are most conducive to the prediction. However, many approaches based on feature importance have limitations.

Among others, the presented features are not necessarily user-friendly in terms of intelligibility. For example, the importance of a single pixel in an image generally conveys little meaningful information [343]. In addition, many feature-importance approaches reach their limits when the number of features is large. The more features there are that are important, the less likely it is that feature-importance attributions will be accessible to users.

To conclude our presentation of explainability approaches, we want to introduce two approaches that could also be considered feature relevance, but do not have the above problems: testing with concept activation vectors (TCAV) and feature visualization (FV).

TCAV TCAV circumvents the above limitations by allowing users to define the features about whose importance they wish to learn. Accordingly, individuals using TCAV do not need to struggle with a myriad of highlighted pixels, but can make targeted queries to the black box. Overall, TCAV is a post-hoc, model-specific (it works only for ANNs), and global approach. For any given so-called *concept*, TCAV measures the extent of that concept's influence on the model's prediction with respect to a particular class. A concept here can be any kind of abstraction, such as a color, an object, or even an idea (e.g., striped, female, etc.).

Because TCAV describes the relationship between a concept and a class, rather than explaining a single prediction, it provides useful information on a model's overall behavior [343]. TCAV is also particularly well-suited to recognize biases in ANNs. To demonstrate this, its inventors used TCAV to find out how important the concepts "male" and "female" are for predicting "doctor" in an image, finding that "male" is significantly more important than "female" [272]. Accordingly, the tested model is biased with respect to gender for doctors.

TCAV is intriguing because it makes use of the latent information that an ANN learned. Normally, an ANN has a fixed feature space (i.e., the set of classes that can be predicted): an ANN trained to distinguish images of cats from images of dogs cannot suddenly output that an image is a rabbit. During training, however, the ANN learns more than just these two classes; in particular, it learns features that can be used to distinguish the classes of interest.

Example #24

Whiskers may be a feature that distinguishes cats from dogs. With TCAV, we can find out whether the concept of "whiskers" is indeed important for predicting "cat" instead of "dog", even though we have never directly tasked the CNN with detecting whiskers.

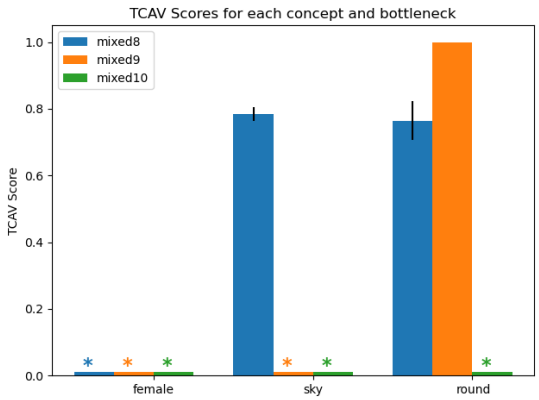


Figure 21: TCAV scores for three different concepts on the prediction "dome".

Figure 21 visualizes the importance of the concepts "female", "sky", and "round" for the class "dome" in specific layers of a CNN. As can be seen, the concept "female" is not important, whereas the concepts "sky" and "round" are, matching our intuitions that domes are round and often photographed against the sky.

FV The final variety of explainability approaches that we wish to mention is FV. FV is only applicable to ANNs classifying data that can be suitably visualized for humans (e.g., pictures or sounds).[78] In other words, it makes sense primarily for CNNs.

(a) Layer 5 (b) Layer 6 (c) Layer 7

Figure 22: FVs of neurons in different layers of a CNN.
The deeper the layer in the network, the more semantically meaningful the recognized things are. While in early layers only abstract edges are recognized, neurons in later layers already seem to recognize things like ropes.

FV aims to make classification in ANNs visible to us, which is accomplished by refining random noise until it fully activates a set of neurons or a single one (e.g., the noise becomes classified as belonging entirely to the class in question) [360, 361]. Thus, the *features* that a neural network decides upon become visualized (and act as the explanation; see Figure 22).

10.2. Criteria for Suitable Explainability Approaches

Given the great variety of approaches, the question arises as to what makes a good one. Unfortunately, there is no consensus on what constitutes a good explainability approach [90, 480, 509]. There are several methods for evaluating explainability approaches, each of which comes with its own underlying rationale for which criteria are essential for good system explainability. However, these evaluation methods all have their limitations. We are not aware of any theoretical considerations that motivate certain quality criteria over others. We will now bridge this gap by distilling philosophically motivated quality criteria that information provided by explainability approaches should meet.

10.2.1. The Three Dimensions of Explainability

We establish, here, three criteria by which to assess the quality of explanatory information produced by an explanatory approach. These criteria are inspired by the three dimensions of explainability described by Baum et al. [62] and Speith [446]:

- **Compehensibility:** Explanatory information must be conveyed in a way that is comprehensible to humans, for humans must understand (certain aspects of) the system based on the information, and comprehensible information is best suited for this purpose.

[78] There are approaches trying to visualize data for all types of ML models, the simplest being *nomograms* (see [310] for early work, see also [347]). These approaches are, however, not very paradigmatic.

- **Fidelity:** Explanatory information must be fidelitous with respect to (the aspect of) the system it is about. For instance, the information must describe the accurate (i.e., correct, real) reasons for a system acting the way it did. Although some details may be omitted or simplified for comprehensibility, outright lies should never be told.

- **Assessability:**[79] Explanatory information must be such that one can assess the satisfaction of a given desideratum. In other words, the quality of explainability depends on the desideratum that one aims to satisfy with it.

While these criteria are inspired by previous work, it is easy to motivate them independently. First, information that is incomprehensible or misleading (because it is not fidelitous) is unlikely to help facilitate a person's true understanding of a phenomenon. Considering the fact that, in the field of explainability, explanations are considered as such only when they lead to understanding (see Section 8.2.1), explainability approaches that produce incomprehensible or infidelitous information are not of high quality and should not be used.

Let us now take a look at assessability. In Section 8.4.1, we argued that the ultimate goal of receiving information about (certain aspects of) a system is to satisfy certain desiderata. However, it is conceptually possible that explanatory information provided by an explainability approach can be both comprehensible and fidelitous, but completely irrelevant to the desiderata of interest. For this reason, we need a criterion that restricts the set of relevant pieces of information to those that also serve the the targeted goal: satisfying these desiderata.

10.2.2. Arguments for Our Criteria

In what follows, we will provide arguments to further corroborate our criteria. In particular, these arguments are based on satisfying certain often cited desiderata (see [105, 292, 294]). Each argument is supported by examples and outlines which criterion it motivates.

Our primary argument is the need for *acceptance* and *trustworthiness*. As argued in the first part of this thesis (see Section 3.1.1 and, specifically, Motivation 1.1), it is plausible to assume that systems that are unable to explain their decisions, predictions, or behavior will lack *acceptance* in the long run. However, the deployment of some kinds of autonomous systems promises to bring about overall positive effects. Thus, as long as people do not accept these systems, their presumably beneficial deployment is threatened [60, 61] (see also Autonomous Vehicle #13).

[79] In the original source [62], the third dimension is *permissibility*. We deliberately deviate here, as we are concerned with the desirable *properties of explanatory information*. The original source dealt with the *properties of computational systems* that are necessary to make them trustworthy. Trustworthy systems should, indeed, not only be be able to deliver assessable information, but they should, in fact, be permissible. That is, they should be positively assessed concerning certain desiderata (e.g., morality, fairness, safety, reliability). However, the information must be assessable with respect to these desiderata to allow somebody to determine whether the system is *de facto* permissible.

Autonomous Vehicle #13

Typical examples that illustrate this argument are autonomous cars. A broad deployment of autonomous cars promises to reduce the number of car accidents, but as studies indicate (e.g., [76]), the kinds of operational autonomous cars that promise to reduce casualties the most would not be trusted today and thus would not gain market share.

For people to accept systems, *trust* is an essential prerequisite [190, 399]. However, not just any kind of trust will be beneficial in the long run. Only *adequate* or *justified* trust in a system will bring about the best consequences concerning it; inadequate trust can lead to disastrous consequences [264]. Consequently, adequate trust is morally desirable.

So, to significantly increase the probability that autonomous systems will be used on a large scale, they must be justifiably trusted. Justified trust is based on trustworthiness [250, 264] (see Motivation 1.1). Furthermore, trustworthiness depends on a stakeholder's justification in believing that a system works properly, and explanations are one way in giving these justifications [264]. Thus, explanations are an important factor in calibrating trust, a connection that is often argued for (see, for example, [70, 222, 434, 457]).

Of course, not just any kind of explanation will do. To be trustworthy, software systems must be able to justify their actions *in the right way*. A well-motivated means of doing so is to provide information that at least meets the criteria above.

First, the information must be comprehensible, because incomprehensible information is not sufficient to provide any justification. Furthermore, the information must be fidelitous, since a lying system, even if otherwise functioning properly, cannot count as trustworthy. Finally, the information must support the assessability of pertinent desiderata, as this allows people to judge whether the system is working properly.[80]

Our second argument is the need for *responsibility* and *autonomy* in the interaction of artificial systems and humans. It is foreseeable that in the future, humans will increasingly depend on decisions by machines. Currently, there is a trend to use ML-based algorithms to make recommendations in morally critical situations (e.g., for health diagnosis).

Example #25

Here is an example for this practice. Avati et al. [45] proposed a DNN for predicting the 3–12 month mortality of a patient based on a certain number of vital parameters recorded over the course of a year. Based on the system's prediction, the patient receives a recommendation of whether or not they should start palliative care.

In such morally critical situations, it is especially desirable that someone can be held responsible or accountable, should something go wrong. However, as we have argued in the

[80] For further discussion of why the quality criteria we set out are good criteria to judge the trustworthiness of an artificial system, see the original introduction of the three dimensions of explainability, [62].

first part of this thesis (see Section 3.1.2 and, specifically, Risk 2.2.2), attributing responsibility is often difficult with modern systems. To still allow for the attribution of responsibility in these situations, it is often assumed that there must be a human in the loop who makes the final decision and is the most probable bearer of responsibility (see Section 9.3).

As became clear in Section 9.3, for that human to properly bear responsibility for a decision, the decision must be made autonomously. The human, however, if not able to *comprehend* and *assess* a recommendation, *loses* autonomy. After all, when someone is to decide competently and autonomously, they need more than just simple recommendations from a system. A human in the loop needs reasons for the recommendations, so as to assess their correctness and potentially challenge them (see Section 9.3.3). Explanatory information, thus, must be at least *comprehensible* and *assessable* to properly allow for human autonomy.

Hiring System #5

Remember Herbert (Section 9.3, e.g., Hiring System #2). He receives a recommendation for each and every applicant, stating whether to keep him or her in the running or not, but nothing else. How can Herbert come to the kind of decision that makes him a bearer of responsibility? If he decides solely in line with the recommendations, he is just a submissive executor of the decision made by an algorithm. In this case, one could simply dismiss the human in the loop altogether. If he decides against the recommendation, he cannot have good reasons for doing so without knowing the reasons for the recommendation in the first place.

The last argument is the need for *fairness*. In many situations, human beings are immediately affected by the decisions made or supported by artificial systems [292, 294].

To ensure that no individual's fundamental rights are violated, a purely statistical checking of the systems is not sufficient. Instead, one needs to be able to assess the *individual* decisions of such systems. In other words, one should not only be concerned with whether a system overall did not discriminate against certain groups of people, but rather with whether each individual decision did not do so. The reason for this is simple: different systems can arrive at the same result in completely dissimilar ways.

Hiring System #6

Imagine that Herbert can choose between two systems that rank job applicants. Both systems rank April last, but for different reasons. The first system does so because she really is the least qualified for the job (e.g., she has no prior work experience, bad grades, etc.). The second system does so just because she is Black and a woman.

While to some this point may seem to be practically unimportant under the assumption that a statistical bias can be reasonably excluded, it is essential to make the system trustworthy

and to establish public acceptance. In particular, it is crucial to guarantee that the fundamental values of liberal societies are considered sufficiently. Statistical adherence to norms and rules is important, but it should not be the only thing to be considered. Each individual case matters. Explanatory information, thus, must be at least *assessable* and *fidelitous*.

10.2.3. Evaluating Contemporary Approaches

Having motivated and argued for our criteria, we will now turn to their application. In particular, we will now look at the exemplary approaches we have introduced to check whether they produce information that meets our criteria. In this regard, it should be noted that we are mostly raising theoretical points, since, as stated above, there is little to no agreement on evaluation methods. Still, this should show that our criteria are worthy of attention. Overall, applying the criteria might reveal the shortcomings of the approaches and, thereby, allow us to create an argument for our argumentation framework, which we will evaluate later.

Comprehensibility Most of the explainability approaches examined benefit from visually representing the explanatory information they produce. One does not have to be an expert to recognize that the highlighted areas of an image correspond to important parts. However, it stands to reason that people without sufficient background in ML will, in many cases, not comprehend what the highlighting means on a deeper (e.g., technical) level [189, 294].

In particular, pixel-oriented saliency masks (produced by, for instance, vanilla backpropagation and integrated gradients) suffer from this. Additionally, these techniques have further problems. First, they generate information that is very difficult to analyze even by experts [13, 189, 294]. For instance, for people without sufficient background, vanilla backpropagation sometimes seems to produce saliency masks with random highlighting (see Figure 10c and Figure 17c). Moreover, as mentioned above, these techniques suffer from computational artifacts, often producing out-of-context highlighting that hinders comprehensibility.

FV also scores rather poorly with respect to comprehensibility, but for different reasons. In particular, neurons responding to human-comprehensible concepts are rare. Thus, the rope-like structures of Figure 22c are the exception, and it is much more common to find something like Figure 22a or Figure 22b. Humans can do rather little with edges as noticed by neurons in early layers. Furthermore, neurons in later layers can respond to multiple higher-level concepts, resulting in strange blended concepts that might be beyond human comprehension.

LIME and TCAV perform better in terms of comprehensibility. When it comes to LIME, its inventors have confirmed the comprehensibility of the information it generates in several studies [393]. TCAV, in contrast, has the advantage that the users themselves can define the concepts for which an ANN is to be tested. On the other hand, TCAV has the disadvantage that it requires at least some background knowledge of how ANNs work, since one must specify which layers of an ANN to test and analyze the results accordingly. For example,

one needs to know that deeper layers of the network are more likely to respond to complex concepts (e.g., gender) than other layers that respond to simpler concepts (e.g., color).

Let us finish this discussion with the evaluation of approaches that do not visualize their information. CIA is understandable because it provides explanatory information that could have come from a human. Every person can clearly think of cases wherein they thought something like "If feature X of this situation were not the case, then I would act differently". Explaining a system's decision in such a counterfactual way seems rather comprehensible.

Fidelity Currently, there is not much research that assesses the fidelity of explanatory information produced by explainability approaches. The reason for this shortcoming is, among others, that it is not entirely clear, formally and technically, what it means for an explainability approach to produce fidelitous information [5, 18, 343].

Despite this difficulty, there are some authors who try to address the issue. For instance, Adebayo et al. [5] developed a sanity check for explainability approaches producing saliency maps. Using this test, they evaluated several well-known approaches. They found that some approaches pass their sanity check and produce fidelitous information (e.g., vanilla backpropagation and Grad-CAM); others do not (e.g., guided backpropagation).

Another notable work is that of Amparore et al. [18]. They found, for instance, that many implementations of LIME do not satisfy the theoretical properties that this approach originally promised. For instance, these implementations produce unstable and infidelitous information. The infidelity of LIME was to be expected, since it is a *model-agnostic* approach. In other words, the internals of the model to be explained are not taken into account in the generated information: it is based solely on an input–output analysis [447].

Continuing with the previous example, plausibly stating that one would have acted otherwise if a certain feature *was not* the case, does not mean that one really acted the way one did because that feature *was* the case. Perhaps the person in the example simply uses the learning excuse as a face-saver to avoid admitting that he or she is not sufficiently intelligent to accomplish the task. Thus, with regard to fidelity, CIA does not yield anything.

Conceivably, this lack of fidelity might be rectified by trying all feasible inputs. Here, the hope could be that, by doing so, one could ascertain that the result in question is accurate (i.e., only exactly the combination of parameters in question led to the result under examination). There are, however, at least two problems with this idea. First, it is extremely likely that ruling out all inputs is practically impossible. Having even one input variable with an infinite domain would cause this. Second, even if it were possible to exclude all other inputs, this is still insufficient because it is very plausible that different information can be provided for the same result. Thus, merely excluding other explanations does not lead to finding the correct one.

Example #26

An excellent example for this problem is the honest merchant as proposed by Immanuel Kant [261]. The merchant in Kant's example is honest because he fears losing his reputation if he is not honest, not because of his duty to be honest. Accordingly, there are at least two explanations for his honesty (see also Hiring System #6).

Correlation (as learned during the training of an ANN) is not necessarily causation. Consequently, CIA most likely only delivers correlations without causal components of the required kind. Causation, however, is both desired and necessary here.

Example #27

Returning to the example used to introduce CIA (Example #23, see also Example #25), in light of the above, it is highly debatable whether there can be a fidelitous CIA-based explanation of the decision that a patient has less than a year to live.

As for the fidelity of FV, this is difficult to assess. However, since the goal of the approach is to completely activate certain neurons, it seems to provide accurate, and thus faithful, information for these neurons. In addition, there are mathematical restrictions that can be incorporated in the FV algorithm so that it displays more robust images, which increases fidelity.

Assessability It should be noted that assessability is a difficult criterion to test because there are so many desiderata that could, in principle, be of interest. For this reason, we will only make some general remarks, starting with CIA and LIME.

With the help of CIA, one can test impermissible configurations of parameters. For example, one can check whether changing the variable "gender" influences a prediction. However, ruling out all impermissible configurations of parameters in CIA is often unfeasible.

In addition to studies on the comprehensibility of the information generated by LIME, Ribeiro et al. have also conducted studies on what participants could do with this information [393]. Among other research outcomes, these studies showed that LIME enables individuals to identify incorrect classifications, and even helps them improve the classifier. This suggests that the provided information is useful for assessing at least some desiderata (e.g., fairness).

Kim et al. compare different approaches that generate saliency maps (viz., vanilla backpropagation, guided backpropagation, integrated gradients, and SmoothGrad) to check whether the information they generate enables one to evaluate whether a given classification makes sense [272]. Their goal is to show that these approaches do not fully suffice.

To this end, they add a visible label, which varies across trials, to the lower left corner of input images. In one trial, the label is constant across individual classes (e.g., all cab images

receive the label "cab"). In another trial, the label sometimes deviates within a class (e.g., some cabs are labeled "cucumber"). In yet another trial, the label is completely random (e.g., each cab image receives an arbitrary label). They found that, in all trials, the image's lower left corner was highlighted by the approaches to a non-negligible extent.

It is important to point out that this does not preclude the approaches from being fidelitous to the classification algorithm, since the lower left corner might actually be used by it (it has a prominent label, after all), even if only insignificantly. However, it does mean that it is difficult to impossible to use the generated information for meaningful assessments.

Coming to the next approaches, Rudin argues that many heatmaps (including those produced by Grad-CAM) do not really allow for good assessments of predictions [404]. The main reason for this is that the heatmap for the most probable class (e.g., wolf) is often hardly distinguishable from that of a less probable class (e.g., flute) [404]. Accordingly, it is questionable whether the obtained information allows for assessing certain desiderata.

Finally, let us talk about TCAV. In our opinion, TCAV performs best when it comes to assessability. This is because it allows for hand-crafted concepts to be reviewed. In other words, a person using TCAV is not simply confronted with an unchanging set of information, as with other approaches, but can inquire after the information that is of interest (e.g., whether gender played a role in the classification, which allows for fairness testing).

Next, we take a look at FV. The scope of FV is fairly restricted, placing no importance on the properties of the feature visualized. Therefore, it is only a coincidence if it can be used to assess a desideratum. Usually, the visualizations reveal nothing of interest.

Table 4: Contemporary explainability approaches evaluated.

Approach	Comprehensibility	Fidelity	Assessability
CIA	+	−	+
LIME	+	−−	+
Vanilla Backpropagation	−−	+	−
Guided Backpropagation	−	−	−
Integrated Gradients	−	?	−
SmoothGrad	+	?	−
Grad-CAM	+	+	−
TCAV	+	+	++
FV	−	+	−−

"+" indicates a positive evaluation of a criterion, and "−" a negative one. "?" indicates that we did not find any data. Note that we took Grad-CAM as prototypical for CAM-based approaches, and counted robustness to CIA.

Summarizing the above, Table 4 offers an overview of how we believe the explanatory information generated by the discussed approaches fare in terms of our criteria. Overall, we think that TCAV fares best, as it is comprehensible, fidelitous, and allows for a wide range of assessments. However, as already mentioned, using TCAV is not necessarily something for laypersons, and the challenges of interpreting its outputs emphasize that there does not seem to be a one-size-fits-all explainability approach.

10.3. Concluding Remarks on the Approaches

As we have demonstrated, except for TCAV, each of the current approaches fails to fulfill at least one of our criteria. We further believe that it is difficult or even impossible to eliminate the approaches' flaws without significantly altering them. Still, we think that they are useful in their own right, and can provide useful insights in some situations, especially for experts.

Comparing these approaches to our framework, one might wonder where the similarities are, and why we spend so much space discussing them. There are several reasons for this. On the one hand, these approaches are a sweep of the field of explainability and, thus, can serve well to give an idea of it. Further, these approaches are largely visual, which gives them intrinsic illustrative potential. In this regard, it should also be mentioned that some of the approaches discussed (e.g., LIME and TCAV) can be used not only visually but also in other modalities, which makes them more versatile than our depiction might suggest.

Additionally, it should be mentioned that all approaches listed here are post-hoc approaches. In other words, they do not intervene in the architecture of a system and change it, but try to explain a system that has already been created. Our framework, however, starts with the architecture of a system and tries to design it in such a way that it is as well suited as possible to generate explanations. While in doing so we do not quite meet the requirement of an ante-hoc explainable system, since the explanations obtained are likely to still be very complex (we will address this in the next section), our framework does also not qualify as a post-hoc approach.

Rather, it may be that our framework needs to be supported by post-hoc explainability approaches. We have already learned about one possible reason for this in Section 6.3.4: if the moral principles have been acquired through an ML process, then we need techniques like the ones above to make them intelligible. Furthermore, if, for example, sub-modules of a system (e.g., the image recognition of a robot) are based on AI, the above approaches can also come into play to make the system's decision making understandable.

With this background now in place, we can finally discuss and evaluate our argumentative framework (Part II of this thesis) based on all the information we collected until now.

Part IV.
Building Bridges

11. Our Framework Revisited

In the last section, we motivated three criteria for evaluating explainability approaches: fidelity, comprehensibility, and assessability, and applied them to evaluate several contemporary approaches. Most of those failed because they do not meet at least one of our criteria (or have other problems). The question that now arises is whether the framework we constructed in Part II of this thesis satisfies the criteria set out.

Before we get into answering this question, we should discuss what kind of explanations we ideally want to extract from our framework.

Medical-Care Robot #30

The medical care robot enters a room with intent to answer a request, but as soon as the robot notices that the patient requires resuscitation, it turns around and leaves, and returns to its recharging station. What happened? Is the robot malfunctioning?

Assuming that the robot has made its decision on the basis of our decision-making method, we could ask an expert to extract an explanation from the logged decision graphs. The explanation we receive could be "The robot decided to go to the recharging station because it realized that the task behind the request it planned to answer was a resuscitation task. It ought to try to resuscitate whenever possible, but its internal representation of its power supply indicated, with an overwhelmingly high probability, that it would not have been able to complete the resuscitation. However, the robot was certain that it could return to the charging station so that it would not run out of power, something it ought to try to avoid whenever possible. Although its duty to try to resuscitate has priority over its duty to avoid running out of power, it came to the overall conclusion that it ought to recharge."

This explanation may make us question the robot's earlier decision. For instance, we could ask why it was going into the room in response to the request at all when its energy level was already so low. The answer we receive could be "The robot originally decided to answer the request because its estimates regarding the tasks behind the request were such that the expected utility of trying to execute the request was higher than the expected utility of going directly to the recharging station. The task most likely associated with the request was fetching water or providing medicine, each of which the robot would have been able to perform and still return to the recharging station afterward. Thus, overall it decided that it ought to execute the request, which came down to the operation of walking to the room."

At this point, we would have sufficient reason to believe that the robot is functioning perfectly well and does not require any repairs. We could justifiably trust it and continue to use it in the context of medical care.

Summarizing our theoretical discussions so far, this example is intended to illustrate that explanations are a crucial factor in our interactions with machines. Only when we receive explanations can we engage appropriately with machines, for example, to develop an adequate trust in them. Explanations are, thus, an essential factor for the acceptance of machines.

Having sketched the type of explanation that we hope to obtain from our framework (although many details must be left for future research), we now examine the three criteria that we set out for suitable explanations. In what follows, we will argue that our approach can, in principle, satisfy them. However, there is some ground to be covered until we can say so.

First, the decision graphs do not come in the nice textual representation we give in Medical-Care Robot #30. While all the information used in these two explanations can be found in the graphs associated with the decisions, only an expert might come to a similar reasoning when looking at the argumentation graphs generated, if at all.

11.1. Comprehensible Explanations through Mapping

With this in mind, the comprehensibility of our proposal seems to be a big issue, and one might wonder how to make the decision graphs of our framework comprehensible to laypersons. This is the first criterion that we tackle. While some parts of the framework are accessible to human comprehension, this does not seem to be the case for others. For these other parts, we propose a concept that we call "mapping" to make them comprehensible.[81]

11.1.1. The Comprehensibility of Our Framework Examined

Let us examine the parts of our framework step by step. First, the restrictions to which the system adheres (i.e., the principles) are comprehensible, because, as we have already mentioned, the principles are most likely based on some laws or societal considerations. These considerations are linked to concepts that can be made available to different audiences. Furthermore, the reasons for introducing a certain principle and not a different one can reasonably be assumed to be saved inside the system. Doing so would yield a function $Perm^{\psi}_{\mathrm{mot}}(\omega, \mathrm{Addressee})$: a function that explains the principle in question accordingly for the stated addressee. Let us explain it via an example considering the health care robot.

Medical-Care Robot #31

There are at least two kinds of possible addressees relevant to the robot: the doctor and the patient, together with his or her relatives.

Assuming that the patient has just been resuscitated, $Perm^{\psi}_{\mathrm{mot}}(\omega, \mathrm{Doctor})$ contains something like this: "The subject's cardiac function was in critical status A-78ZB. Regulation BS-Z200-9 required me to try to resuscitate him." Only persons with specific training such as doctors can understand such a sentence.

[81] Unfortunately, we will only be able to treat this concept superficially, and we conceptualize it as a matter to be developed in future research.

Consequently, $Perm^{\psi}_{\text{mot}}(\omega, \text{Patient})$ differs from the equivalent oriented towards the doctor. Now, this function would output something like "You had a heart attack. Consequently, I tried to resuscitate you." Here, we have a sentence that could later be used when the robot attempts to justify its actions.

In the explanations of Medical-Care Robot #31, the robot made no gibberish statements for the respective addressee, nor used terms that the addressee would not know. The explanations are comprehensible for the corresponding recipients, and they are plausibly both encoded in the same graph. Overall, it can be assumed that rational humans would come to the same decision as the system, given the same data and a similar reasoning approach.

This is also why we opted for an argumentative approach in our framework (see Section 6.1). Arguments are a rational tool that humans use to come to decisions. Therefore, if a decision-making process is based on an argumentative process (as described in Section 6.2) this process is, in principle, suitable for being comprehensible for a human being.

For this reason, we believe that it is justified to see promising grounds for *rationalizing* explanation in the argumentation graphs. Recall that our goal has been to establish a close connection between a system's deliberative processes and its explanations. With this approach, we aim to live up to Davidson's idea: by appealing to the system's information (i.e., its knowledge and the principles) in the explanations, its outputs and behavior can be rationalized.

However, as mentioned above, the decision graphs resulting from our framework seem to be too complex for laypersons to understand. In particular, the means to obtain comprehensible textual representations as in Medical-Care Robot #30 and Medical-Care Robot #31 is unclear. To pave the way, we will examine the concept of *mapping* explanations in what follows.

11.1.2. Mapping Explanations

To explain what we mean by "mapping", we have to take a step back and examine the work of Erasmus et al. [171]. They argue that, when it comes to the three accounts of scientific explanation that we have previously discussed (i.e., DNE, CME, NME), all medical AI systems are explainable in a certain sense.[82] Their concept of explainability, however, differs from the one that we are using in this thesis. For them, understanding as the pragmatic success condition for explanations is not what constitutes an explanation. They believe that a set of statements qualifying as an explanation depends merely on its form and the relation between the individual statements. Taking this into account, computational systems are explainable by means of, for example, DNEs. This is the case because such systems are traditionally based on certain regularities that can take on the role of laws in the explanation. Thus, we can form DNEs to explain the systems. Furthermore, for other types of scientific explanation the same also holds true, and even complex systems based on DL are explainable in this way [171].

[82] They also argue the same for another type of explanation, namely *inductive-statistical explanations*.

Now, despite their focus on the form of explanation, Erasmus et al. also hold that the concept of understanding is of central importance, especially in the medical domain. To this end, they draw a distinction between explainability and interpretability. For Erasmus et al., the process of converting or translating a possibly complex explanation into a more comprehensible one is the interpretation (see Quote 6). Now, while the verb "to interpret" can be used in this way, that usage deviates from the use we have outlined above. However, since we find their idea of translating possibly incomprehensible explanations into comprehensible ones to be helpful for our purposes, we will make use of it, but refer to it as "mapping" instead.

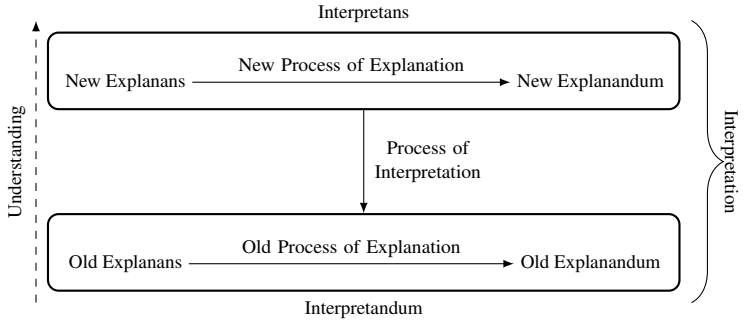


Figure 23: Interpretation (i.e., mapping) according to Erasmus et al. [171].

Erasmus et al. draw several different pictures of how such a mapping may come to be. The most general form is depicted in Figure 23. In this case, all constituents of an explanation (i.e., explanans, explanandum, and process of explanation) are replaced by alternative ones. The new explanation is used to interpret the old one, facilitating its understanding. This is what we call "mapping".

As one can imagine, mapping explanations is interesting in our case, where we want to make the complex explanations extracted from our argumentation graph comprehensible. However, there are some problems. First and foremost, as all constituents of the explanation are replaced during mapping, fidelity is at risk. Here, one implicit assumption is that the explananda bear a certain similarity, and that the new explanation is directed at the same phenomenon. We will come back to this later (in Section 11.2).

The concept of mapping brings us back to the distinction we made between interpretability and explainability. Although we do not adopt Erasmus et al.'s usage here, their proposal has some interesting implications that we will explore to revisit notions of interpretability, explainability, and mapping, in order to connect these concepts.

11.1.3. Revisiting Interpretability

In the last section, we introduced the concept of mapping as a possible way to generate comprehensible explanations from our argumentation graph. While we have to leave the technical details of such a mapping to future research, we will illuminate some theoretical considerations linked to it in this section. In particular, we will outline how the concept of mapping is linked to how we envision the distinction between explainability and interpretability.

Let us first recap the distinction. For us, interpretability is concerned with associating human-comprehensible concepts to the processes occurring in artificial systems. Explainability, on the other hand, is concerned with communicating these concepts to different kinds of human recipients, so as to evoke their understanding of the original processes.

But how exactly are interpretability, mapping, and explainability related, then? Basically, our idea is that the internal workings of a system can naturally be described by scientific explanations (e.g., DNEs). Such explanations would be consistent with Dennett's physical stance. Through an initial mapping to the same or a different kind of scientific explanation, the inner workings are associated with concepts that are comprehensible for humans. This would be "interpretability". Afterwards, *explainability* comes into play via another mapping process. The potentially still-complex scientific explanations are translated into another form of explanation (e.g., reason explanations) that is more comprehensible to laypersons.

In terms of our framework, interpretability is trivial because the principles that underlie the reasoning process are human-comprehensible (as described above). Moreover, the algorithm by which the system arrives at its decision is deterministic, such that it has clearly defined rules that can be used to form DNEs. What is left is the translation to a reason explanation. Here, the $Perm^{\psi}_{\mathrm{mot}}(\omega, \mathrm{Addressee})$ function is a first step. However, there is the additional need for efficient heuristics that help to tone down the argumentation graph's complexity, such that it does not overwhelm most addressees anymore. We leave such heuristics to future research.

At this point, we would like to briefly anticipate the discussion on assessability. Conceiving interpretability, mapping, and explainability in the way just described, we can make a perfect connection to the satisfaction of desiderata. To this end, let us take trustworthiness as an example. Kästner et al. [264] suggested trustworthiness to have two conditions: the system has to be competent, and an addressee has to justifiedly believe that it is so.

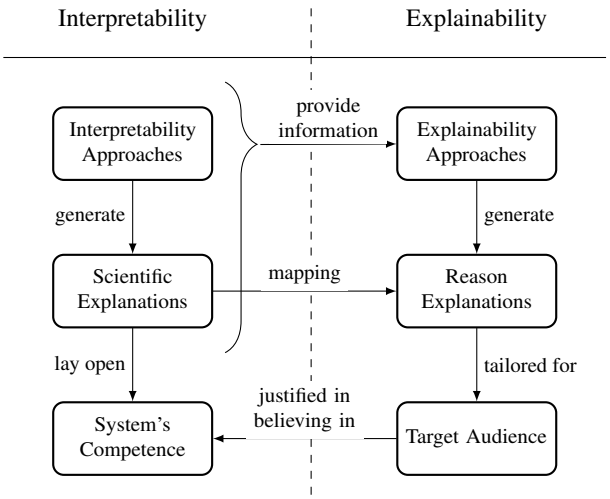


Figure 24: The relationship between interpretability and explainability.

Now, the DNEs that our framework generates are a perfect way to lay open a system's competence. In principle, they describe the system's behavior with complete fidelity, neither missing nor withholding anything. Furthermore, the mapped reason explanations are human-comprehensible, providing an addressee with the perfect justification for believing in the system's competence, should it indeed be competent. With this, our approach (supported by an adequate mapping mechanism) is a fitting way not only to assess, but even to facilitate, the system's trustworthiness. The whole process above is visualized in Figure 24.

11.2. The Other Criteria Examined

After tackling comprehensibility, we will now come to fidelity and assessability. Since these two are interlinked, we will deal with them in the same section. First, our demand for fidelity is, in principle, satisfied by the decision-making process of our framework, since the decisions are truly based on the argumentative process. In other words, there are no unknown, black-box components in the evaluative algorithm of our approach.

Medical-Care Robot #32

By assumption, the robot's explanations were generated and extracted from the graphs in the robot's logs. Furthermore, by assumption, these graphs are the graphs that led to the robot's decisions (rather than being constructed afterwards). Thus, they are based on the true deliberations of the robot and, consequently, are fidelitous.

However, fidelity suffers from the mapping process. As described above, it is likely that we will need to map explanations to ensure comprehensibility, sacrificing some fidelity in the process: the mapped explanations will inevitably omit some details or simplify facts in order to be comprehensible. Still, we will argue that this is not a problem, at least from a pragmatic point of view. To this end, we will take a look at another form of mapping proposed by Erasmus et al. and its relation to limiting-case relationships in the philosophy of science.

11.2.1. Fidelity and Limiting-Case Relationships

Thinking back to the most general form of mapping (see Figure 23), all constituents of an explanation (i.e., explanans, explanandum, and process of explanation) are replaced. This obviously causes fidelity issues, as not even the explanandum is the same.

However, it is often not required to do such a radical form of mapping. In particular, to ensure a minimal degree of fidelity, the explanandum should remain fixed (see Figure 25). As we will outline, this mapping process is similar to limiting-case relationships in science.

A limiting case occurs when the predictions of a scientific theory can, under certain limiting or boundary conditions, be estimated by using a less complex theory. The Newtonian laws of motion, for instance, are approximately true in cases where we are not dealing with velocity

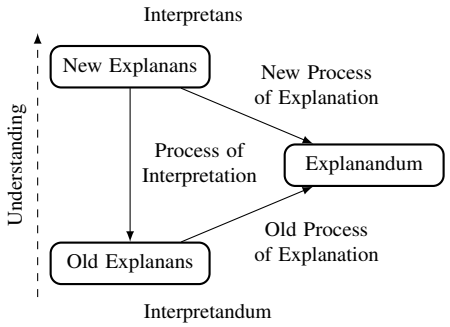


Figure 25: A more specific form of interpretation (i.e., mapping) [171].

near the speed of light. For this reason, the Newtonian laws are a limiting case of the relativity theory. Even simpler, Galilei's law of falling body is a limiting case of the Newtonian laws when it comes to falling processes near the earth's surface (see Digression #7).

Digression #7

In order to infer Galilei's law of falling bodies from the Newtonian laws, we need Newton's gravitational law and his so-called "second law of motion".

Gravitational law Between any two (ideally) spherical bodies of masses m_1 and m_2 there is a mutual gravitational force, whose magnitude F is given by the equation

$$F = \gamma \frac{m_1 \cdot m_2}{\mathrm{dist}(m_1, m_2)^2} \tag{2}$$

Here $\mathrm{dist}(m_1, m_2)$ is the distance between the centers of the bodies with the masses m_1 and m_2, and γ is a constant of nature, the gravitational constant, with the approximate value of $6.67 \cdot 10^{-11}$ $\mathrm{m^3 kg^{-1} s^{-2}}$.

Newton's second law of motion If a force of magnitude F acts on a body of mass m, and thereby causes an acceleration of magnitude a in this body in the direction of the acting force, then the relation between force, mass, and acceleration is given by the equation $F = m \cdot a$.

The boundary case holds when we look at falling processes near the earth's surface. We want to calculate how much the earth's gravitational force accelerates a falling object. In order to do so, we equate the force described by Newton's gravitational law with the force in Newton's second law in of motion order to calculate the acceleration a. We write m_{object} for the falling object's mass, and m_{earth} for the earth's mass.

$$m_{\text{object}} \cdot a_{\text{object}} = \gamma \frac{m_{\text{object}} \cdot m_{\text{earth}}}{\text{dist}(m_{\text{object}}, m_{\text{earth}})^2} \tag{3}$$

Now we can cancel m_{object} from the left and from the fraction.

$$a_{\text{object}} = \gamma \frac{m_{\text{earth}}}{dist(m_{\text{object}}, m_{\text{earth}})^2} \tag{4}$$

Finally, we insert the $5.9722 \cdot 10^{24}$ kg for the earth's mass and $6,371,000$ m (the earth's radius) for the distance between the two objects. Here we make, taken strictly, the mistakes: we do not factor in the falling object's changing distance from the earth's surface. However, as we are concerned with falling processes near the earth's surface, this mistake is negligible.

$$a = \frac{6.67 \text{ m}^3}{10^{11} \text{ kg s}^2} \cdot \frac{5.972210^{24} \text{ kg}}{(6,371,000 \text{ m})^2} \approx 9.814 \frac{\text{m}}{\text{s}^2} \tag{5}$$

Overall, we arrive at an acceleration of roughly 9.814 ms^2, which is the constant g in Galilei's law of falling bodies: $d(t) = \frac{1}{2} \cdot g \cdot t^2$

Why does a limiting case qualify as a mapping process? Let us illustrate the similarities by means of an example. Using a DNE, we can explain the same phenomenon (e.g., the falling of a ball) by reference to at least three different scientific theories and their respective laws: by reference to Galilei's law of falling bodies, by reference to Newtonian mechanics, and by reference to general relativity. Although Galilei's law of falling bodies is, strictly speaking, false,[83] it is more comprehensible for laypeople than the other theories.

Through this digression on boundary case relations, we can illustrate why mapping has pragmatic advantages that make up for the decreased fidelity. First, Galilei's law of falling bodies is indubitably useful for making reliable predictions in many everyday situations. Furthermore, it is still taught in schools today, even though there are theories that are far superior. This goes so far that the intricacies of general relativity are taught, if at all, only in high school classes or advanced courses. Hence, Galilei's theory suffices for most everyday situations and makes vital parts of the falling process comprehensible for different recipients.

Likewise, one could argue that the infidelity we accept in return for a more comprehensible argumentative process is justifiable, given the pragmatic benefits involved. As we have argued, satisfying desiderata is the overarching goal of most explainability endeavors, and understanding the system is a necessary condition for doing so. Accordingly, only if the "thought process" of a system can be made comprehensible for all participants, can explanations fulfill their desiderata. This being said, we will briefly discuss a way to make the results more comprehensible, while retaining fidelity, before coming to assessability.

[83] Most likely, all three of these theories are, strictly speaking, false, but this is outside of the scope of this thesis.

11.2.2. Fidelity and Different Output Formats

While the mapping just discussed requires some arguments to still be considered fidelitous, there is another type of mapping proposed by Erasmus et al. that does not require such arguments. This last type of mapping solely changes the process of explanation; the explanans and explanandum remain the same (see Figure 26). Against this background, questions now arise as to what this might look like in terms of the explanations produced by explainability approaches and, further, why the mapping can still be considered fidelitous.

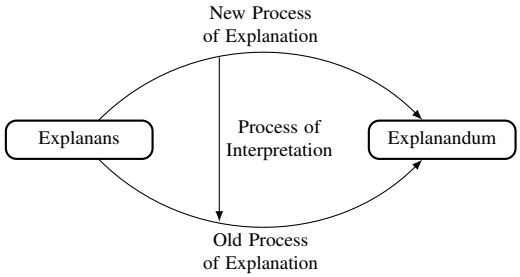


Figure 26: The final type of interpretation (i.e., mapping) [171].

Concerning the former question, one possibility for such a mapping could be the use of different output formats. It is easy to see that presenting the same explanandum in different ways can facilitate or impede understanding, depending on the situation and the recipient. In situations that require quick decisions, an overly complex representation (e.g., a long text with many instructions) could overwhelm a person and, thus, hinder comprehension, whereas simple visual representations could have the opposite effect. Given recipients with a lot of time on their hands, however, textual representations might be better at facilitating understanding of more complex issues.

This type of mapping is interesting because it is frequently found in the explainability debate. In particular, many explainability approaches allow for different forms of representation. Often these are gradient-based, which usually have a numerical output. These numerical values are then used to create heat maps for visualization (see Section 10.1.3).

As for the latter question, we believe that the form of representation does not change the represented information as such. One can represent one and the same proposition in both numerical and visual form without (significantly) changing its content. For this reason, this form of representation does not affect the explanation's fidelity.

11.2.3. Assessability

The final criterion is assessability. In our framework, the explanations we obtain are, in principle, assessable for many desiderata. Since a representation of the world is embedded in our framework, we can simulate new restrictions or circumstances before they are tested in real situations. Therefore, we can assess whether the system behaves properly, given specific restrictions or specific circumstances even before deployment.

Medical-Care Robot #33

By assumption, the explanations delivered by the robot are based on the principles built into the robot. The robot explicitly invokes them to justify its actions, which is one of the crucial aspects of assessability. Another aspect is that the robot makes explicit where instrumental reasoning was involved (i.e., in the expected utility statements).

Finally, we could, in principle, change the knowledge of the system—either directly or by simulation—and compute the resulting graph offline in order to determine what the robot would have decided, and for what reasons, in these alternate situations. Having the ability to change the robot's knowledge enables *counterfactual checking*, and is much more than can ever be hoped for in assessing human beings—about whose behavior only rough estimates can be made.

Let us come back to the link between fidelity and assessability. One motivation for fidelity as a quality criterion for explanations is that the assessment we make, based on those explanations, is correct. While we could mislead people by telling them how smoothly the system works, this claim would be a blatant lie, were the system flawed. Accordingly, it would undermine trustworthiness and acceptance, as stated above (see Section 10.2.2).

Now, as long as fidelity is ensured with respect to the desideratum one wants to assess, fidelity with respect to other aspects is no longer that paramount. For this reason, slightly infidelitous explanations, such as are likely to result from the mapping process, are unproblematic in certain cases. Such explanations can still contribute to the goal of explainability, and thus help with understanding the system to satisfy desiderata.

Our approach has been shown able to satisfy all three criteria, although we will have to leave some details to future research (e.g., how exactly the mapping should be done). Thus, our approach seems tailor-made to provide machine explanations. However, how well it performs in real implementations needs to be evaluated in future research, too.

Unfortunately, the chances that our approach will find wide range application in the near future seem slim. There is an increasing use of ML when it comes to the (moral) decision making of artificial systems. It is easy to create systems in this way, and unfortunately, economic considerations for development of systems are often the major driving force for the development rather than any ethical considerations. Thus, from an economic point of view, our approach would have to be much more advanced to have a real competitive advantage. However, as scientists, we will always strive for an ideal solution which is socially and ethically sound, and will promote that and improve it until it comes to be in use.

12. Conclusion

In this thesis, we argued that machine ethics and machine explainability are necessary to augment each other. The view that they are mutually reinforcing is not as widespread as we feel it should be. Only in symbiosis can machine ethics and explainability achieve their full potential.

12.1. Summary

Regarding machine ethics, we have argued that to align machines morally is currently more pressing than to enable complete moral behavior. With respect to machine explainability, we have outlined the field by differentiating some terms and presenting a model. In this regard, we have also ordered the field per se, identifying that understanding is required to satisfy desiderata. These desiderata, in turn, trace back to machine ethics.

As a practical connection point, we introduced a formal and general framework combining machine ethics and machine explainability. The motivation of the framework was to provide a method of morally constrained decision making that, at the same time, enables explanation. To this end, we proposed an approach to decision making that is based on arguments. By applying three criteria—comprehensibility, fidelity, and assessibility—we argued that our approach is promising. As we have identified this approach as promising, we can envisage many more aspects that should be subjects of future research.

12.2. Future Research

In our discussions, several details were deferred for future work. Regarding machine ethics, a more thorough discussion of its advantages and potential disadvantages in light of the concept of moral alignment seems a worthwhile idea.

For our framework, numerous questions remained unanswered. We did not address several optimization issues. For example, there could be significantly fewer world states to consider if some variables that make up ω are dependent. Also, we have ignored the fact that some variables may have very large, continuous, or even infinite ranges, so considering all possible cases could be practically infeasible or even impossible. Efficient heuristics are needed here to limit the number of options to the most likely or important ones.

Additionally, there are at least five interesting and pressing interdisciplinary research questions that remain open:

- There is the question of how to model the content and ordering of principles in a more sophisticated manner, as well as how to quantify these orderings—and whether this is even necessary. After all, in light of our results (especially those found in Section 7.1), one could be inclined to switch to a framework that relies on something other than

principles. Exploring the ideas of Dietrich [148], who bases decision making on reason, could be an interesting avenue of further research in this case.

- The principle order might be context-dependent. This would essentially necessitate particularism rather than generalism, particularism being the belief that there are *no* general principles governing what ought to be done and that, instead, normative reasons vary from context to context in an unsystematic manner [134, 135].
- Decisions must be made regarding the question of how to aggregate and weigh reasons wherein the answer might well depend on the context of the application.
- We have also postponed the question of how to handle cases involving *epistemic* uncertainty (i.e., pure non-determinism) to future research as well.
- Finally, we have left open exactly how the mapping of explanations should be done. To find an answer to this question, researchers from computer science, philosophy, and psychology should come together.

Coming to machine explainability, the most interesting tasks are further analysis and ordering of the field. In this work we have taken first steps in this direction, but there is much more to do because the current research landscape is so jumbled. In addition, it is imperative to get a better understanding of the desiderata and, in particular, how machine explainability is supposed to satisfy them. Machine explainability has, as we have shown, enormous untapped potential to positively contribute to society—a little more research can make a huge difference here.

ML is becoming increasingly important for the implementation of artificial systems. In order to avoid dystopian situations where we no longer understand the actions of the machines that surround us, we need to be able to explain artificial systems and even black-box systems. This will be a difficult, but, in our opinion, not impossible, undertaking. Consequently, there is certainly more than enough work to be done in terms of machine ethics and the machine explainability.

Appendices

A. Code Listings

LIME

```
# Fix threading problems on macOS
import os
os.environ['KMP_DUPLICATE_LIB_OK']='True'

# Read image
from skimage import io
from skimage import transform
Xi = io.imread("dome.jpg")
Xi = transform.resize(Xi, (299,299))
Xi = (Xi - 0.5)*2   # Pre-processing for InceptionV3
io.imshow(Xi/2+0.5) # Show image before pre-processing
io.show()

# Predict class for image using InceptionV3
import numpy as np
import keras
from keras.applications.imagenet_utils import decode_predictions
np.random.seed(222)
inceptionV3_model = keras.applications.inception_v3.InceptionV3()
preds = inceptionV3_model.predict(Xi[np.newaxis,:,:,:])
top_pred_classes = preds[0].argsort()[-5:][::-1]
print(decode_predictions(preds)[0]) # Print top 5 classes

# Generate segmentation for image
from skimage import segmentation
superpixels = segmentation.quickshift(Xi, kernel_size=4,max_dist=200,
    ratio=0.2)
num_superpixels = np.unique(superpixels).shape[0]
io.imshow(segmentation.mark_boundaries(Xi/2+0.5, superpixels))
io.show()

# Generate perturbations
num_perturb = 150
perturbations = np.random.binomial(1, 0.5, size=(num_perturb,
    num_superpixels))

# Create function to apply perturbations to images
import copy
def perturb_image(img,perturbation,segments):
    active_pixels = np.where(perturbation == 1)[0]
    mask = np.zeros(segments.shape)
    for active in active_pixels:
        mask[segments == active] = 1
```

```
    perturbed_image = copy.deepcopy(img)
    perturbed_image = perturbed_image*mask[:,:,np.newaxis]
    return perturbed_image

# Show examples of perturbations
io.imshow(perturb_image(Xi/2+0.5,perturbations[0],superpixels))
io.show()
io.imshow(perturb_image(Xi/2+0.5,perturbations[1],superpixels))
io.show()
io.imshow(perturb_image(Xi/2+0.5,perturbations[2],superpixels))
io.show()

predictions = []
for pert in perturbations:
    perturbed_img = perturb_image(Xi,pert,superpixels)
    pred = inceptionV3_model.predict(perturbed_img[np.newaxis,:,:,:])
    predictions.append(pred)

predictions = np.array(predictions)

# Compute distances to original image
from sklearn import metrics
original_image = np.ones(num_superpixels)[np.newaxis,:]
distances = metrics.pairwise_distances(perturbations,original_image,
    metric='cosine').ravel()

# Transform distances to a value between 0 an 1 using a kernel function
kernel_width = 0.25
weights = np.sqrt(np.exp(-(distances**2)/kernel_width**2))
print(weights.shape)

# Estimate linear model
from sklearn.linear_model import LinearRegression
class_to_explain = top_pred_classes[0]
simpler_model = LinearRegression()
simpler_model.fit(X=perturbations, y=predictions[:,:,class_to_explain],
    sample_weight=weights)
coeff = simpler_model.coef_[0]

# Use coefficients from linear model to extract top features
num_top_features = 3
top_features = np.argsort(coeff)[-num_top_features:]

# Show only the superpixels corresponding to the top features
mask = np.zeros(num_superpixels)
mask[top_features]= True # Activate top superpixels
io.imshow(perturb_image(Xi/2+0.5,mask,superpixels))
```

```
io.show()
```

Listing 1: LIME implemented in Python.

Saliency Maps

```
import tensorflow as tf
import numpy as np

from matplotlib import pylab as plt
from matplotlib.transforms import Bbox

import PIL.Image

import saliency.core as saliency


extent = 0, 300, 0, 300
bbox_inches = Bbox([[1.02, 0.26], [5.56, 4.48]])


def showImage(img, name='image'):
    plt.imshow(img, extent=extent)
    plt.savefig('results/' + name + '.png', bbox_inches=bbox_inches)
    plt.show()


def showAttributionImage(img, name='attribution'):
    plt.imshow(img, cmap='inferno', vmin=0, vmax=1, extent=extent)
    plt.savefig('results/' + name + '.png', bbox_inches=bbox_inches)
    plt.show()


def showSuperimposedImage(img1, img2, name='superimposed'):
    plt.imshow(img1, extent=extent)
    plt.imshow(img2, cmap='inferno', alpha=0.5, extent=extent)
    plt.savefig('results/' + name + '.png', bbox_inches=bbox_inches)
    plt.show()


def showHeatMap(img, name='heatmap'):
    plt.imshow(img, cmap='inferno', extent=extent)
    plt.savefig('results/' + name + '.png', bbox_inches=bbox_inches)
    plt.show()


def loadImage(file_path):
```

```
    img = PIL.Image.open(file_path)
    img = img.resize((299, 299))
    img = np.asarray(img)
    return img


def preprocessImage(img):
    img = tf.keras.applications.inception_v3.preprocess_input(img)
    return img


def loadModel():
    return tf.keras.applications.inception_v3.InceptionV3()


m = loadModel()
conv_layer = m.get_layer('mixed10')
model = tf.keras.models.Model([m.inputs], [conv_layer.output, m.output])

class_idx_str = 'class_idx_str'


def call_model_function(images, call_model_args=None, expected_keys=None):
    target_class_idx = call_model_args[class_idx_str]
    images = tf.convert_to_tensor(images)
    with tf.GradientTape() as tape:
        if expected_keys == [saliency.base.INPUT_OUTPUT_GRADIENTS]:
            tape.watch(images)
            _, output_layer = model(images)
            output_layer = output_layer[:, target_class_idx]
            gradients = np.array(tape.gradient(output_layer, images))
            return {saliency.base.INPUT_OUTPUT_GRADIENTS: gradients}
        else:
            conv_layer, output_layer = model(images)
            gradients = np.array(tape.gradient(output_layer, conv_layer))
            return {saliency.base.CONVOLUTION_LAYER_VALUES: conv_layer,
                    saliency.base.CONVOLUTION_OUTPUT_GRADIENTS: gradients
    }


# Load the image
im_orig = loadImage('./dome.jpg')
im = preprocessImage(im_orig)

# Show the image
showImage(im_orig, name='original')
```

```
87
88 _, predictions = model(np.array([im]))
89 prediction_class = np.argmax(predictions[0])
90 call_model_args = {class_idx_str: prediction_class}
91
92 print("Prediction class: " + str(prediction_class))
93
94 def vanillaGradients():
95     # Construct the saliency object. This alone doesn't do anthing.
96     gradient_saliency = saliency.GradientSaliency()
97
98     # Compute the vanilla mask and the smoothed mask.
99     vanilla_mask_3d = gradient_saliency.GetMask(im, call_model_function,
    call_model_args)
100     smoothgrad_mask_3d = gradient_saliency.GetSmoothedMask(im,
    call_model_function, call_model_args)
101
102     # Call the visualization methods to convert the 3D tensors to 2D
    grayscale.
103     vanilla_mask_grayscale = saliency.VisualizeImageGrayscale(
    vanilla_mask_3d)
104     smoothgrad_mask_grayscale = saliency.VisualizeImageGrayscale(
    smoothgrad_mask_3d)
105
106     # Render the saliency masks.
107     showAttributionImage(vanilla_mask_grayscale, name='
    vanillaGradientsMask')
108     showAttributionImage(smoothgrad_mask_grayscale, name='
    smoothGradVanillaMask')
109     showSuperimposedImage(im_orig, vanilla_mask_grayscale, name='
    vanillaGradients')
110     showSuperimposedImage(im_orig, smoothgrad_mask_grayscale, name='
    smoothGradVanilla')
111
112
113 def integratedGradients():
114     # Construct the saliency object. This alone doesn't do anthing.
115     integrated_gradients = saliency.IntegratedGradients()
116
117     # Baseline is a black image.
118     baseline = np.zeros(im.shape)
119
120     # Compute the vanilla mask and the smoothed mask.
121     vanilla_integrated_gradients_mask_3d = integrated_gradients.GetMask(
122         im, call_model_function, call_model_args, x_steps=25, x_baseline=
    baseline, batch_size=20)
```

```
    # Smoothed mask for integrated gradients will take a while since we
    are doing nsamples * nsamples computations.
    smoothgrad_integrated_gradients_mask_3d = integrated_gradients.
    GetSmoothedMask(
        im, call_model_function, call_model_args, x_steps=25, x_baseline=
    baseline, batch_size=20)

    # Call the visualization methods to convert the 3D tensors to 2D
    grayscale.
    vanilla_mask_grayscale = saliency.VisualizeImageGrayscale(
    vanilla_integrated_gradients_mask_3d)
    smoothgrad_mask_grayscale = saliency.VisualizeImageGrayscale(
    smoothgrad_integrated_gradients_mask_3d)

    # Render the saliency masks.
    showAttributionImage(vanilla_mask_grayscale, name='
    integratedGradientsMask')
    showAttributionImage(smoothgrad_mask_grayscale, name='
    smoothGradIntegratedMask')
    showSuperimposedImage(im_orig, vanilla_mask_grayscale, name='
    integratedGradients')
    showSuperimposedImage(im_orig, smoothgrad_mask_grayscale, name='
    smoothGradIntegrated')


vanillaGradients()
integratedGradients()
```

Listing 2: Different saliency methods implemented in Python.

Guided Backpropagation

```
import os

import tensorflow as tf
from matplotlib.transforms import Bbox
from tensorflow.keras.applications.inception_v3 import InceptionV3,
    preprocess_input
import tensorflow.keras.backend as kb
from tensorflow.keras.models import Model
from tensorflow.keras.preprocessing import image
import numpy as np
import matplotlib.pyplot as plt

# Fix threading problems on macOS
os.environ['KMP_DUPLICATE_LIB_OK'] = 'True'

```

```
15
16 def build_model():
17     return InceptionV3()
18
19
20 # Load and preprocess image
21 def load_image(path):
22     x = image.load_img(path, target_size=(299, 299))
23     x = image.img_to_array(x)
24     x = np.expand_dims(x, axis=0)
25     x = preprocess_input(x)
26     return x
27
28
29 def deprocess_image(x):
30     # normalize tensor: center on 0., ensure std is 0.25
31     x = x.copy()
32     x -= x.mean()
33     x /= (x.std() + kb.epsilon())
34     x *= 0.25
35
36     # clip to [0, 1]
37     x += 0.5
38     x = np.clip(x, 0, 1)
39
40     # convert to RGB array
41     x *= 255
42     if kb.image_data_format() == 'channels_first':
43         x = x.transpose((1, 2, 0))
44     x = np.clip(x, 0, 255).astype('uint8')
45     return x
46
47
48 @tf.RegisterGradient("GuidedRelu")
49 def _GuidedReluGrad(op, grad):
50     gate_f = tf.cast(op.outputs[0] > 0, "float32")  # for f^l > 0
51     gate_R = tf.cast(grad > 0, "float32")  # for R^l+1 > 0
52     return gate_f * gate_R * grad
53
54
55 @tf.custom_gradient
56 def guidedRelu(x):
57     def grad(dy):
58         return tf.cast(dy > 0, "float32") * tf.cast(x > 0, "float32") *
    dy
59
60     return tf.nn.relu(x), grad
```

```
61
62
63 # process example input
64 preprocessed_input = load_image("dome.jpg")
65
66 model = build_model()
67 gb_model = Model(inputs=[model.inputs], outputs=[model.get_layer('mixed10
       ').output])
68 layer_dict = [layer for layer in gb_model.layers[1:] if hasattr(layer, '
       activation')]
69 for layer in layer_dict:
70     if layer.activation == tf.keras.activations.relu:
71         layer.activation = guidedRelu
72
73 with tf.GradientTape() as tape:
74     inputs = tf.cast(preprocessed_input, tf.float32)
75     tape.watch(inputs)
76     outputs = gb_model(inputs)
77
78 extent = 0, 300, 0, 300
79 bbox_inches = Bbox([[1.02, 0.26], [5.56, 4.48]])
80
81 grads = tape.gradient(outputs, inputs)[0]
82
83 plt.imshow(np.flip(deprocess_image(np.array(grads)), -1), extent=extent)
84 plt.savefig("results/guidedBackPropagationMask.png", bbox_inches=
       bbox_inches)
85 plt.show()
86
87 plt.imshow(preprocessed_input[0], extent=extent)
88 plt.imshow(np.flip(deprocess_image(np.array(grads)), -1), alpha=0.75,
       extent=extent)
89 plt.savefig("results/guidedBackPropagation.png", bbox_inches=bbox_inches)
90 plt.show()
91
92 img = preprocessed_input[0]
93
94 stdev = 0.15 * (np.max(img) - np.min(img))
95 h, w = 299, 299
96 arr = np.zeros((h, w, 3), np.float)
97 N = 25
98
99 for x in range(N):
100     noise = np.random.normal(0, stdev, img.shape).astype(np.float32)
101     img_plus_noise = img + noise
102     with tf.GradientTape() as tape:
103         tmp_inputs = tf.cast([img_plus_noise], tf.float32)
```

```
        tape.watch(tmp_inputs)
        tmp_outputs = gb_model(tmp_inputs)
    tmp_grads = tape.gradient(tmp_outputs, tmp_inputs)[0]
    tmp_img = np.flip(deprocess_image(np.array(tmp_grads)), -1)
    img_arr = np.array(tmp_img, dtype=np.float)
    arr = arr + img_arr / N

arr = np.array(np.round(arr), dtype=np.uint8)
plt.imshow(arr, extent=extent)
plt.savefig("results/smoothGuidedBackPropagationMask.png", bbox_inches=
    bbox_inches)
plt.show()

plt.imshow(img, extent=extent)
plt.imshow(arr, alpha=0.75, extent=extent)
plt.savefig("results/smoothGuidedBackPropagation.png", bbox_inches=
    bbox_inches)
plt.show()
```

Listing 3: Guided backpropagation implemented in Python.

CAM

```
from tensorflow.keras.preprocessing.image import load_img
from tensorflow.keras.applications.inception_v3 import InceptionV3,
    preprocess_input, decode_predictions
import matplotlib.pyplot as plt
import cv2
import numpy as np
from matplotlib.transforms import Bbox
from gradcamutils import GradCam, GradCamPlusPlus, ScoreCam,
    build_guided_model, GuidedBackPropagation, superimpose, \
     read_and_preprocess_img


def build_model():
    return InceptionV3(include_top=True, weights='imagenet')


model = build_model()
layer_name = 'mixed10'

img_path = 'dome.jpg'
orig_img = np.array(load_img(img_path), dtype=np.uint8)
img_array = read_and_preprocess_img(img_path, size=(299, 299))

predictions = model.predict(img_array)
```

```
top = decode_predictions(predictions, top=5)[0]
print(img_path)
print("class activation map for:", top[0])

grad_cam = GradCam(model, img_array, layer_name)
grad_cam_superimposed = superimpose(img_path, grad_cam)
grad_cam_emphasized = superimpose(img_path, grad_cam, emphasize=True)

grad_cam_plus_plus = GradCamPlusPlus(model, img_array, layer_name)
grad_cam_plus_plus_superimposed = superimpose(img_path,
    grad_cam_plus_plus)
grad_cam_plus_plus_emphasized = superimpose(img_path, grad_cam_plus_plus,
     emphasize=True)

score_cam = ScoreCam(model, img_array, layer_name)
score_cam_superimposed = superimpose(img_path, score_cam)
score_cam_emphasized = superimpose(img_path, score_cam, emphasize=True)

faster_score_cam = ScoreCam(model, img_array, layer_name, max_N=10)
faster_score_cam_superimposed = superimpose(img_path, faster_score_cam)
faster_score_cam_emphasized = superimpose(img_path, faster_score_cam,
    emphasize=True)

guided_model = build_guided_model(build_model)
saliency = GuidedBackPropagation(guided_model, img_array, layer_name)
saliency_resized = cv2.resize(saliency, (orig_img.shape[1], orig_img.
    shape[0]))

grad_cam_resized = cv2.resize(grad_cam, (orig_img.shape[1], orig_img.
    shape[0]))
guided_grad_cam = saliency_resized * grad_cam_resized[..., np.newaxis]

grad_cam_plus_plus_resized = cv2.resize(grad_cam_plus_plus, (orig_img.
    shape[1], orig_img.shape[0]))
guided_grad_cam_plus_plus = saliency_resized * grad_cam_plus_plus_resized
    [..., np.newaxis]

score_cam_resized = cv2.resize(score_cam, (orig_img.shape[1], orig_img.
    shape[0]))
guided_score_cam = saliency_resized * score_cam_resized[..., np.newaxis]

faster_score_cam_resized = cv2.resize(score_cam, (orig_img.shape[1],
    orig_img.shape[0]))
guided_faster_score_cam = saliency_resized * faster_score_cam_resized
    [..., np.newaxis]

img_gray = cv2.imread(img_path, 0)
```

```
dx = cv2.Sobel(img_gray, cv2.CV_64F, 1, 0, ksize=3)
dy = cv2.Sobel(img_gray, cv2.CV_64F, 0, 1, ksize=3)
grad = np.sqrt(dx ** 2 + dy ** 2)
grad = cv2.dilate(grad, kernel=np.ones((5, 5)), iterations=1)
grad -= np.min(grad)
grad /= np.max(grad)  # scale 0. to 1.

grad_times_grad_cam = grad * grad_cam_resized
grad_times_grad_cam_plus_plus = grad * grad_cam_plus_plus_resized
grad_times_score_cam = grad * score_cam_resized
grad_times_faster_score_cam = grad * faster_score_cam_resized

extent = 0, 300, 0, 300
bbox_inches = Bbox([[1.02, 0.26], [5.56, 4.48]])


def showImage(img, name='image'):
    plt.imshow(img, extent=extent)
    plt.savefig('results/' + name + '.png', bbox_inches=bbox_inches)
    plt.show()


showImage(orig_img, name="original")

showImage(grad_cam, name="gradCamMask")
showImage(grad_cam_plus_plus, name="gradCamPlusPlusMask")
showImage(score_cam, name="scoreCamMask")
showImage(faster_score_cam, name="fasterScoreCamMask")

showImage(grad_cam_superimposed, name="gradCam")
showImage(grad_cam_plus_plus_superimposed, name="gradCamPlusPlus")
showImage(score_cam_superimposed, name="scoreCam")
showImage(faster_score_cam_superimposed, name="fasterScoreCam")

showImage(grad_cam_emphasized, name="gradCamEmphasized")
showImage(grad_cam_plus_plus_emphasized, name="gradCamPlusPlusEmphasized"
    )
showImage(score_cam_emphasized, name="scoreCamEmphasized")
showImage(faster_score_cam_emphasized, name="fasterScoreCamEmphasized")
```

Listing 4: Different CAM methods implemented in Python

TCAV

```
import numpy as np
import keras
from keras.models import load_model
```

```
4  from keras.models import model_from_json
5  import tcav.activation_generator as act_gen
6  import tcav.cav as cav
7  import tcav.model as tcav_model
8  import tcav.tcav as tcav
9  import tcav.utils as utils
10 import tcav.utils_plot as utils_plot  # utils_plot requires matplotlib
11 import os
12 import tensorflow as tf
13 from skimage import io
14 from skimage import transform
15 from keras.applications.imagenet_utils import decode_predictions
16
17 # Fix threading problems on macOS
18 os.environ['KMP_DUPLICATE_LIB_OK'] = 'True'
19
20 # The directories
21 working_dir = './tcav_class_test/'
22 activation_dir = working_dir + '/activations/'
23 cav_dir = working_dir + '/cavs/'
24 source_dir = "./image_net_subsets/"
25
26 bottlenecks = ['mixed8', 'mixed9', 'mixed10']
27
28 utils.make_dir_if_not_exists(activation_dir)
29 utils.make_dir_if_not_exists(working_dir)
30 utils.make_dir_if_not_exists(cav_dir)
31
32 # this is a regularizer penalty parameter for linear classifier to get
      CAVs.
33 alphas = [0.1]
34
35 target = 'dome'
36 concepts = ['female', 'sky', 'round']
37
38 sess = utils.create_session()
39 model = keras.applications.inception_v3.InceptionV3()
40 model.summary()
41
42
43 def read_image(file_name):
44     img = io.imread(file_name)
45     img = transform.resize(img, (299, 299))
46     img = (img - 0.5)*2
47     return img
48
49
```

```
50 def read_directory(directory):
51     images = []
52     for file_name in os.listdir(directory):
53         if file_name.endswith(".jpg"):
54             images.append(read_image(directory + file_name))
55         else:
56             continue
57     return images
58 
59 
60 images = read_directory(source_dir + target + "/")
61 print(len(images))
62 for img in images:
63     preds = model.predict(img[np.newaxis, :, :, :])
64     preds = decode_predictions(preds)[0]
65     if preds[0][1] != 'dome' or preds[0][2] < 0.75:
66         print(preds)
67         io.imshow(img/2+0.5)
68         io.show()
69 
70 print('ready')
71 
72 
73 # Modified version of PublicImageModelWrapper in TCAV's models.py
74 # This class takes a session which contains the already loaded graph.
75 # This model also assumes softmax is used with categorical crossentropy.
76 class CustomPublicImageModelWrapper(tcav_model.ImageModelWrapper):
77     def __init__(self, sess, labels, image_shape,
78                  endpoints_dict, name, image_value_range):
79         super(self.__class__, self).__init__(image_shape)
80 
81         self.sess = sess
82         self.labels = labels
83         self.model_name = name
84         self.image_value_range = image_value_range
85 
86         # get endpoint tensors
87         self.ends = {'input': endpoints_dict['input_tensor'], 'prediction
    ': endpoints_dict['prediction_tensor']}
88 
89         self.bottlenecks_tensors = self.get_bottleneck_tensors()
90 
91         # load the graph from the backend
92         graph = tf.compat.v1.get_default_graph()
93 
94         # Construct gradient ops.
95         with graph.as_default():
```

```
96            self.y_input = tf.compat.v1.placeholder(tf.int64, shape=[None
    ])
97
98            self.pred = tf.expand_dims(self.ends['prediction'][0], 0)
99            self.loss = tf.reduce_mean(
100               tf.compat.v1.nn.softmax_cross_entropy_with_logits_v2(
101                   labels=tf.one_hot(
102                       self.y_input,
103                       self.ends['prediction'].get_shape().as_list()[1])
    ,
104                   logits=self.pred))
105       self._make_gradient_tensors()
106
107   def id_to_label(self, idx):
108       return self.labels[idx]
109
110   def label_to_id(self, label):
111       return self.labels.index(label)
112
113   @staticmethod
114   def create_input(t_input, image_value_range):
115       """Create input tensor."""
116       def forget_xy(t):
117           """Forget sizes of dimensions [1, 2] of a 4d tensor."""
118           zero = tf.identity(0)
119           return t[:, zero:, zero:, :]
120
121       t_prep_input = t_input
122       if len(t_prep_input.shape) == 3:
123           t_prep_input = tf.expand_dims(t_prep_input, 0)
124       t_prep_input = forget_xy(t_prep_input)
125       lo, hi = image_value_range
126       t_prep_input = lo + t_prep_input * (hi-lo)
127       return t_input, t_prep_input
128
129   @staticmethod
130   def get_bottleneck_tensors():
131       """Add Inception bottlenecks and their pre-Relu versions to
    endpoints dict."""
132       graph = tf.compat.v1.get_default_graph()
133       bn_endpoints = {}
134       for op in graph.get_operations():
135           if 'ConcatV2' in op.type:
136               name = op.name.split('/')[0]
137               bn_endpoints[name] = op.outputs[0]
138
139       return bn_endpoints
```

```

# input is the first tensor, logit and prediction is the final tensor.
# note that in keras, these arguments should be exactly the same for other models (e.g VGG16), except for the model name
endpoints_v3 = dict(
    input=model.inputs[0].name,
    input_tensor=model.inputs[0],
    logit=model.outputs[0].name,
    prediction=model.outputs[0].name,
    prediction_tensor=model.outputs[0],
)

# instance of model wrapper, change the labels and other arguments to whatever you need
labels = [OMITTED]
mymodel = CustomPublicImageModelWrapper(sess, labels, [299, 299, 3], endpoints_v3, 'Inception_V3', (-1, 1))

act_generator = act_gen.ImageActivationGenerator(mymodel, source_dir, activation_dir, max_examples=100)

tf.compat.v1.logging.set_verbosity(0)

num_random_exp = 3  # folders (random500_0, random500_1)
mytcav = tcav.TCAV(sess, target, concepts, bottlenecks, act_generator, alphas,
                   cav_dir=cav_dir, num_random_exp=num_random_exp)

results = mytcav.run(run_parallel=False)

utils_plot.plot_results(results, num_random_exp=num_random_exp)
```

Listing 5: TCAV implemented in Python.

FV

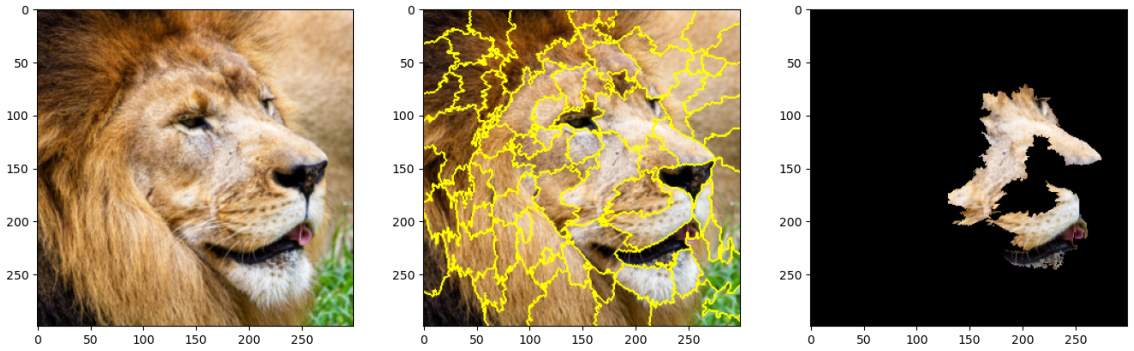
```
import numpy as np
import os
import tensorflow as tf
assert tf.__version__.startswith('1')

import lucid.modelzoo.vision_models as models
from lucid.misc.io import show
import lucid.optvis.objectives as objectives
import lucid.optvis.param as param
import lucid.optvis.render as render
```

```
import lucid.optvis.transform as transform
import matplotlib.pyplot as plt

#InceptionV3/InceptionV3/Conv2d_1a_3x3/Relu
#InceptionV3/InceptionV3/Conv2d_2a_3x3/Relu
#InceptionV3/InceptionV3/Conv2d_2b_3x3/Relu
#InceptionV3/InceptionV3/Conv2d_3b_1x1/Relu
#InceptionV3/InceptionV3/Conv2d_4a_3x3/Relu
#InceptionV3/InceptionV3/Mixed_5b/concat
#InceptionV3/InceptionV3/Mixed_5c/concat
#InceptionV3/InceptionV3/Mixed_5d/concat
#InceptionV3/InceptionV3/Mixed_6a/concat
#InceptionV3/InceptionV3/Mixed_6b/concat
#InceptionV3/InceptionV3/Mixed_6c/concat
#InceptionV3/InceptionV3/Mixed_6d/concat
#InceptionV3/InceptionV3/Mixed_6e/concat
#InceptionV3/InceptionV3/Mixed_7a/concat
#InceptionV3/InceptionV3/Mixed_7b/concat
#InceptionV3/InceptionV3/Mixed_7c/concat
#InceptionV3/Predictions/Softmax

#model = models.InceptionV3_slim()
model = models.InceptionV1()
model.load_graphdef()

path = './test/'
os.makedirs(path, exist_ok=True)

# Test 1
#param_f = lambda: param.image(128, batch=2)
#obj = objectives.channel("mixed4a_pre_relu", 492, batch=1) - objectives.
    channel("mixed4a_pre_relu", 492, batch=0)
#img = render.render_vis(model, obj, param_f)
#tf.keras.preprocessing.image.save_img(path + "test11.png", img[0][0])
#tf.keras.preprocessing.image.save_img(path + "test12.png", img[0][1])

# Test 2
#param_f = lambda: param.image(128, batch=4)
#obj = objectives.channel("mixed4a_pre_relu", 97) - 1e2*objectives.
    diversity("mixed4a")
#img = render.render_vis(model, obj, param_f)
#tf.keras.preprocessing.image.save_img(path + "test21.png", img[0][0])
#tf.keras.preprocessing.image.save_img(path + "test22.png", img[0][1])
#tf.keras.preprocessing.image.save_img(path + "test23.png", img[0][2])
#tf.keras.preprocessing.image.save_img(path + "test24.png", img[0][3])

# Test 3
```

```
neuron1 = ('mixed4b_pre_relu', 111)      # large fluffy
# neuron1 = ('mixed3a_pre_relu', 139)    # pointilist
# neuron1 = ('mixed3b_pre_relu',  81)    # brush strokes
# neuron1 = ('mixed4a_pre_relu',  97)    # wavy
# neuron1 = ('mixed4a_pre_relu',  41)    # frames
# neuron1 = ('mixed4a_pre_relu', 479)    # B/W

neuron2 = ('mixed4a_pre_relu', 476)      # art
# neuron2 = ('mixed4b_pre_relu', 360)    # lattices
# neuron2 = ('mixed4b_pre_relu', 482)    # arcs
# neuron2 = ('mixed4c_pre_relu', 440)    # small fluffy
# neuron2 = ('mixed4d_pre_relu', 479)    # bird beaks
# neuron2 = ('mixed4e_pre_relu', 718)    # shoulders

C = lambda neuron: objectives.channel(*neuron)

img1 = render.render_vis(model, C(neuron1))
img2 = render.render_vis(model, C(neuron2))
img3 = render.render_vis(model, C(neuron1) + C(neuron2))
tf.keras.preprocessing.image.save_img(path + "test31.png", img1[0][0])
tf.keras.preprocessing.image.save_img(path + "test32.png", img2[0][0])
tf.keras.preprocessing.image.save_img(path + "test33.png", img3[0][0])

#transforms = [
#        transform.pad(16),
#        transform.jitter(8),
#        transform.random_scale([n/100. for n in range(80, 120)]),
#        transform.random_rotate(list(range(-10, 10)) + list(range(-5, 5)
    ) + 10*list(range(-2, 2))),
#        transform.jitter(2)
#]

#param_f = lambda: param.image(128, fft=True, decorrelate=True)

#for i in range(5):
#        obj = objectives.channel("InceptionV3/InceptionV3/Mixed_4a/
    concat", i)
#        img = render.render_vis(model, obj, param_f, transforms=
    transforms)
#        img = render.render_vis(model, obj)
#        tf.keras.preprocessing.image.save_img(path + "neuron" + str(i) +
     ".png", img[0][0])
```

Listing 6: FV implemented in Python.

B. Further Images

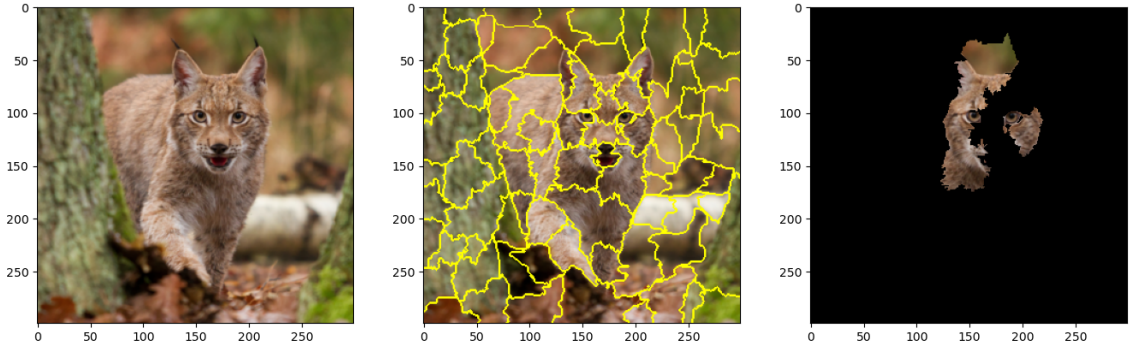


(a) Original image with prediction “lion”.

(b) Image segmented into interpretable clusters.

(c) Most relevant clusters for the prediction.

Figure 27: LIME visualized for the image of a lion.

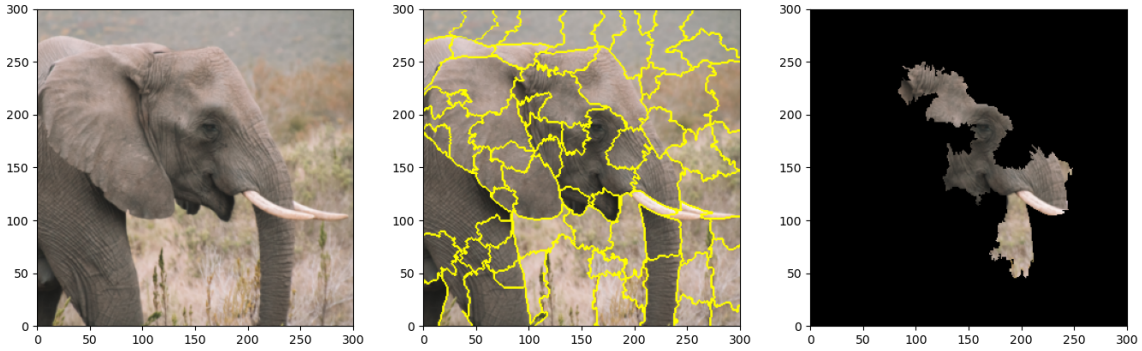


(a) Original image with prediction “lynx”.

(b) Image segmented into interpretable clusters.

(c) Most relevant clusters for the prediction.

Figure 28: LIME visualized for the image of a lynx.

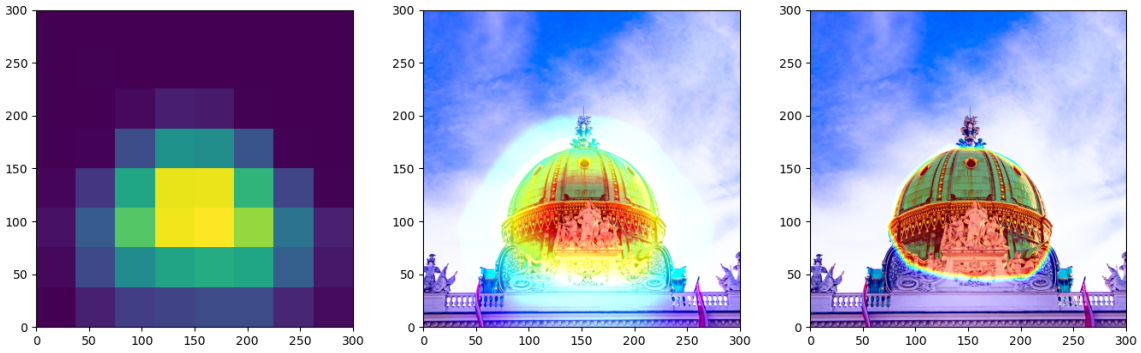


(a) Original image with prediction “elephant”.

(b) Image segmented into interpretable clusters.

(c) Most relevant clusters for the prediction.

Figure 29: LIME visualized for the image of an elephant.

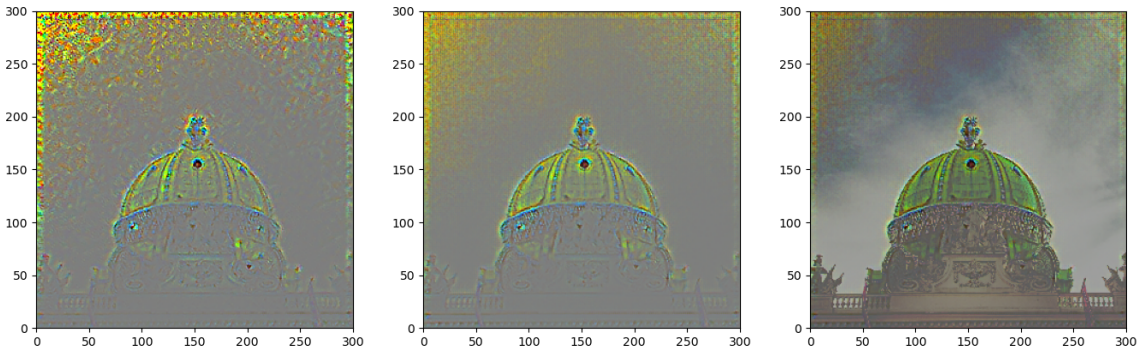


(a) Original feature activation mask.

(b) The activation mask superimposed.

(c) The activation mask condensed.

Figure 30: More images for Grad-CAM.



(a) Guided backpropagation mask.

(b) SmoothGrad of guided backpropagation mask.

(c) SmoothGrad of guided backpropagation.

Figure 31: More images for guided backpropagation.

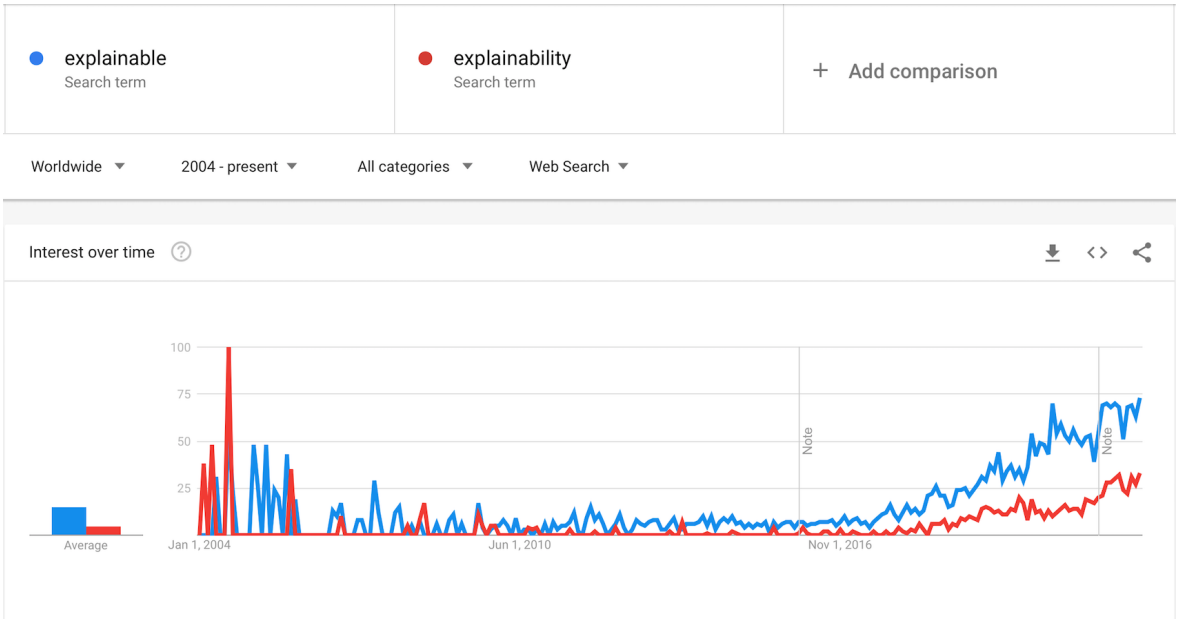


Figure 32: Google Trends search for the terms “explainable” (blue) and “explainability” (red). The large spikes around 2004 are likely due to inaccurate measurement methods and, thus, to be disregarded.

C. Literature Review

To find out more about explainability, its facets and implications, we conducted a systematic literature review together with Larissa Chazette and Wasja Brunotte. The procedure of this review is described in [105] and [106]. In order to be able to use the results of this literature review as our own contribution in this work, we did not adopt them one-to-one. Rather, we have taken intermediate results of the literature review, and processed them in such a way that they fit better into the context of this thesis. Thus, the results mentioned here should constitute enough of our own work to count as an independent contribution.

More precisely, our own work starts right after the initial coding (see [105] for prior steps). Starting from the initial coding, we completely re-coded all the extracted data from the more than 200 papers. In other words, Larissa Chazette and Wasja Brunotte helped to select the papers, and they also helped to extract relevant parts from these papers (initial coding), but we have completely revised the use of these parts. Specifically, we used these parts to extract information on the relationship between explainability and its related concepts (e.g., transparency, interpretability) on the one hand, and desiderata on the other. Often, in the process, we revisited the original articles to see if there were more interesting information to extract. Overall, this re-coding was necessary for our purposes because the extracted information was originally used to draw conclusions in the context of requirements engineering. However, since this work is concerned with desiderata that support a link from machine explainability to machine ethics, the focus of re-coding was on such desiderata.

All desiderata obtained by this, together with the sources supporting a link between machine explainability and them, can be found in Table 5. Furthermore, all extracted citations from which we read such a connection can be found in Table 6–Table 39. We also published a subset of this research in an earlier version in [294]. There, additional information is added on whether the claimed connection was supported by empirical investigations. Unfortunately, due to the scope of the investigation in this thesis, we were unable to take this step again. In addition, we would like to emphasize that we have extracted the quotes to the best of our abilities. However, spelling errors cannot be ruled out due to the sheer volume of sources.

Table 5: Sources for all desiderata.

Desideratum	Sources
Acceptance	[29, 43, 48, 51, 70, 73, 103, 110, 111, 116, 123, 124, 127, 137, 152, 165, 166, 173, 179, 190, 199, 222, 234, 274, 280, 304, 305, 348, 358, 372, 373, 382, 383, 386, 388, 394, 395, 397, 399, 410, 412, 422, 465, 478, 488, 490, 501, 502, 510]
Accountability	[1, 72, 95, 139, 156, 175, 189, 287, 302, 303, 311, 322, 341, 345, 349, 365, 388, 391, 395, 397, 425–427, 442, 443, 452, 476, 485, 490]
Accuracy	[4, 29, 51, 73, 95, 118, 123, 131, 137, 179, 187, 203, 236, 274, 277, 280, 287, 303, 307, 322, 372, 389, 399, 425, 427, 465, 478, 508]
Autonomy	[30, 175, 177, 222, 342, 373, 388, 397]

Continued on next page ...

Table 5 – continued from previous page

Quote	Sources
Confidence	[30, 51, 54, 73, 101, 123, 127, 150, 165–167, 173, 177, 179, 180, 199, 222, 274, 304, 348, 350, 358, 391, 422, 444, 464–466, 473, 479, 493, 501, 508, 510]
Controllability	[1, 4, 30, 96, 112, 203, 222, 339, 342, 348, 358, 388, 397, 399, 417, 426, 465, 473]
Debuggability	[4, 29, 30, 73, 95, 99, 101, 103, 131, 137, 139, 150, 165, 168, 189, 219, 221, 257, 267, 276, 277, 282, 283, 304, 305, 312, 322, 337, 341, 342, 345, 358, 391, 399, 400, 425–427, 441–443, 476, 485, 488, 490, 492, 497, 510]
Education	[30, 108, 110, 111, 130, 137, 139, 199, 222, 225, 231, 281, 282, 308, 325, 339, 341, 345, 348, 357, 358, 386, 388, 426, 444, 471, 479, 492, 493]
Effectiveness	[43, 51, 72, 82, 110, 111, 150, 165, 180, 199, 222, 229, 348, 358, 382, 385, 386, 452, 463–466, 478, 485, 489, 505, 508, 510]
Efficiency	[4, 30, 51, 72, 82, 108, 110, 111, 113, 137, 151, 156, 229, 287, 291, 322, 348, 358, 366, 382, 425, 427, 463–466, 489]
Fairness	[1, 4, 6, 54, 72, 92, 95, 99, 123, 139, 150, 155, 175, 188, 225, 227, 233, 236, 259, 274, 280, 302, 304, 322, 337, 342, 358, 388, 391, 397, 399, 417, 441–443, 472, 475, 481, 485, 503, 506]
Informed Consent	[173, 342, 373]
Legal Compliance	[1, 4, 30, 43, 48, 72, 103, 113, 127, 150, 165, 173, 177, 179, 189, 193, 201, 225, 231, 233, 241, 251, 277, 280, 302, 303, 322, 341, 342, 345, 349, 365, 386, 391, 394, 399, 427, 436, 441, 442, 476, 488, 490, 503, 506]
Morality	[1, 54, 72, 155, 165, 175, 189, 234, 236, 304, 322, 342, 373, 397, 399, 417, 488]
Performance	[30, 43, 54, 95, 108, 123, 124, 137, 173, 199, 222, 231, 236, 254, 274, 277, 304, 349, 366, 388, 389, 391, 394, 397, 427, 436, 442, 472, 476, 481, 482, 503, 510]
Persuasiveness	[51, 70, 72, 110, 111, 124, 137, 165, 180, 199, 229, 304, 339, 348, 358, 394, 399, 410, 417, 426, 463–466, 487, 490]
Privacy	[54, 99, 112, 127, 155, 189, 233, 236, 241, 287, 391, 417, 442, 443, 510, 512]
Reliability	[1, 93, 99, 155, 189, 357, 362, 389, 475, 482]
Reliance	[72, 93, 231, 388]
Responsibility	[95, 130, 156, 175, 339, 349, 358, 365, 373, 388, 397, 426]
Robustness	[54, 73, 77, 99, 155, 187, 322]
Safety	[29, 30, 99, 155, 187, 201, 214, 236, 280, 288, 350, 373, 391, 399, 443, 490]
Satisfaction	[51, 70, 72–74, 96, 110, 111, 137, 151, 152, 166, 173, 198, 199, 225, 229, 275, 291, 305, 348, 358, 388, 422, 436, 463–466, 473, 478, 487]
Science	[4, 99, 155, 179, 234, 280, 304, 308, 394, 397, 399, 417, 441, 442, 454, 476, 485, 488, 497, 503]
Security	[128, 236, 241, 277, 373, 399, 442, 443, 479, 503]
Transferability	[54, 109, 234, 417, 497]
Transparency	[1, 29, 30, 43, 51, 70, 72, 108, 110, 111, 118, 137, 150, 165, 177, 189, 199, 215, 222, 229, 234, 235, 241, 291, 302, 322, 326, 348, 358, 365, 373, 382, 383, 386, 394, 426, 427, 442, 444, 464–466, 478, 479, 485, 503, 511]
Trust	[1, 4, 30, 43, 48, 51, 69, 72, 73, 77, 82, 93, 95, 96, 99, 101, 103, 110–112, 116, 118, 123, 124, 127, 128, 131, 137, 150, 151, 155, 156, 161, 165, 167, 173, 175, 177, 179, 180, 188, 190, 196, 198, 199, 201, 214, 215, 221, 222, 229, 231, 233–236, 241, 274, 275, 277, 280, 288, 289, 304, 305, 312, 316, 322, 337, 339, 341, 342, 348–350, 357, 358, 362, 365, 366, 373, 382, 383, 386, 389, 391, 393–395, 399, 410, 412, 415, 417, 422, 426, 427, 441, 442, 451, 456, 463–466, 478, 479, 485, 487–490, 492, 503, 505, 506, 508, 510, 511]
Trustworthiness	[30, 54, 130, 137, 180, 187, 231, 274, 277, 326, 341, 393, 397]
Understandability	[1, 4, 30, 43, 48, 51, 54, 71, 77, 93, 95, 96, 99, 108, 113, 116, 123, 124, 127, 131, 150, 151, 161, 165–167, 175, 177, 180, 182, 193, 203, 209, 219, 221, 222, 231, 233–235, 254, 274, 277, 282, 303, 312, 322, 325, 326, 339, 358, 372, 388, 389, 391, 394, 395, 399, 410, 412, 414, 422, 425–427, 441, 444, 451, 452, 454, 465, 476, 482, 485, 490, 493, 505, 508, 510, 511]
Usability	[51, 70, 108, 112, 137, 155, 199, 201, 231, 236, 254, 277, 304, 358, 388, 427, 463–466, 473, 511]
Usefulness	[70, 111, 165, 180, 198, 199, 221, 254, 382, 388, 410, 465, 502]
Validation	[71, 99, 137, 156, 164, 165, 188, 189, 325, 365, 489]
Verification	[71, 93, 123, 137, 173, 241, 322, 325, 341, 342, 350, 358, 427, 444, 490, 493]

Table 6: Quotes for the desideratum *Acceptance*.

Quote	Src.
Intelligent systems that are explaining their decisions to increase the user's [...] acceptance are widely studied.	[29]
[...] appropriate approaches and methods require [...] appropriate explanation-aware techniques e.g. for increasing the acceptance of the patterns and their evaluation [...].	[43]
[...] different aspects should be taken into account while producing an explanation in order to increase user acceptance.	[48]
In addition to improving user acceptance of recommendations [...], explanations can serve a multiplicity of aims [...].	[51]
A system's ability to explain its recommendations in a way that makes its reasoning more transparent can contribute significantly to users' acceptance of its suggestions.	[70]
[...] systems' ability to explain their reasoning has been found to be critical to users' acceptance of their decisions [...].	[70]
Explanation has been shown to be important for user acceptance [...] in a number of studies.	[73]
[...] explanations in general and justifications in particular make the generated advice more acceptable to users [...].	[73]
The ability to generate explanations holds the key [...] towards acceptance of AI-based systems [...].	[103]
The importance of explanation interfaces in providing system transparency and thus increasing user acceptance has been recognized in a number of fields [...].	[110]
The importance of explanation interfaces in [...] increasing user acceptance has been recognized in a number of fields [...].	[111]
[...] explanation can cause users to overestimate item quality, which may [...] stop users from using the system again [...].	[111]
XAI will be key for both expert and non-expert users to enable them to have a deeper understanding and the appropriate level of trust, which will hopefully lead to increased adoption of this vital technology.	[116]
If explanations can increase transparency and interpretability, this might improve [...] acceptance from both students and educators [...]	[386]
Systems researchers have emphasized the importance of explanations as a means of influencing user acceptance [...] in systems by increasing confidence in systems' abilities [...].	[123]
[...] explaining to the user why a recommendation was made increased acceptance of the recommendations.	[124]
Increasing transparency of user-adaptive systems could thus increase [...] acceptance of such systems [...].	[124]
[...] explanations that are too complex might actually negatively affect acceptance of a system.	[124]
[...] self-driving cars, which can demonstrate transparency in operations, will help promote trust, which is pivotal to its acceptance by society [...].	[127]
The perceptions arising from the use of explanation facilities include [...] user acceptance [...].	[137]
[...] explanation can enhance the acceptability of expert systems.	[137]
A common finding in most of these studies show that the inclusion of explanation helps improve user-acceptance in expert systems [...].	[137]
[...] the inclusion of justification explanations had a profound impact on user acceptance of the system [...].	[137]
[...] user acceptance is very high when explanations are used and provided [...].	[137]
[...] benefit [of] explanation facilities [...] include greater user acceptance of the recommender system as a decision making aid.	[137]
Explanation [...] can substantially affect [...] acceptance [...].	[137]
A common finding in most of these studies show that the inclusion of explanation helps improve user-acceptance in expert systems, although the conclusions in a small number of studies run counter to these findings.	[137]
Explaining automatic recommendations [...] has shown an important effect on users' acceptance over the items recommended.	[152]
[...] studies [...] often focus on how explanations can improve acceptance of recommender systems [...].	[165]
[...] previous research has established that explanations help users to [...] accept a recommendation.	[165]
AI rationalization has a number of potential benefits over other explainability techniques: [...] humanlike communication [...] may afford [...] advantages such as higher degrees of [...] willingness to use autonomous systems [...].	[166]
[...] they could show that transparency increased the acceptance of the recommendations.	[173]
[...] the user requires good explanations from the system as a requirement for model acceptance [...].	[179]
[...] explanation systems can address [...] trust concerns [...] and thus can help to move [...] one step closer to [...] acceptance by end users [...].	[190]

Continued on next page ...

Table 6 – continued from previous page

Quote	Src.
Explanations, by virtue of making the performance of a system transparent to its users, are influential for user acceptance of intelligent systems [. . .].	[199]
Explanations that conform to Toulmin's model should be more persuasive [. . .]. Thus, they should lead to greater [. . .] acceptance.	[199]
[. . .] providing explanations can improve the acceptance of ACF systems [. . .].	[222]
Some of the benefits provided [by explanation facilities] are: [. . .] Acceptance. Greater acceptance of the recommender system as a decision aide [. . .].	[222]
We believe that by providing transparency into the workings of the ACF process, we will [. . .] increase [users'] willingness to use the ACF system as a decision aid.	[222]
The result [from adding explanations] will be filtering systems that are more accepted [. . .].	[222]
[. . .] transparency and understanding of the AI systems' behavior is inevitably, e.g., to increase user acceptance.	[234]
Another experiment found 'why' explanations to increase recommendation acceptance [. . .]	[274]
To the extent that public acceptance of ML algorithms requires that end users have some grasp of the inner workings [. . .], the notion of interpretation acquires heightened importance.	[280]
"Explanation is often embraced as a cure for ""black box"" models to gain trust and adoption."	[304]
[. . .] we discuss how [. . .] providing explanations [. . .] has the potential to increase user satisfaction and thus acceptance [. . .].	[305]
[. . .] generating explanations of application behavior [. . .] has been employed [. . .] with the goal of increasing [. . .] acceptance of these systems.	[305]
The importance of [. . .] explanation on improving user satisfaction (e.g., acceptance, trust) has been extensively discussed.	[348]
How to visualize explanation about recommendations is important for user acceptance of recommender systems.	[348]
[. . .] Some benefits provided by explaining recommendations such as: [. . .] acceptance.	[348]
However, to be accepted by end users, the suggestions [. . .] must be perceived to be fair and transparent [. . .], and explanations are key to this.	[358]
Explaining decisions returned by intelligent systems is [. . .] essential for gaining acceptance [. . .].	[372]
Lack of such explanations does not prevent users from being able to operate the devices, but may nevertheless make them refrain from doing so.	[373]
The importance of explanation interfaces in providing system transparency and thus increasing user acceptance has been well recognized in a number of fields [. . .].	[382]
The importance of explanation interfaces in providing system transparency and thus increasing user acceptance has been well recognized in a number of fields [. . .].	[383]
Explanations [. . .] promote [. . .] acceptance of the system [. . .].	[388]
[. . .] in many, if not most, cases, the explanation is beneficial to the system's acceptance [. . .].	[394]
The ability to help people understand their decisions through explanations [. . .] will [. . .] make them more willing to continue the use of AI systems.	[395]
With an explanation of the algorithm's decision, it is possible for human beings to accept, disregard, challenge, or overrule that decision.	[397]
[. . .] the agent might need to provide information about its decision to help convince the human participant of the correctness of their solution, aiding in the adoption of these agent based technologies [. . .].	[399]
In both cases, the information the agent provides should build trust to ensure its decisions are accepted [. . .].	[399]
To date, many reasons have been suggested for making systems explainable [. . .]: [. . .] To justify its decisions so the human participant can decide to accept them [. . .].	[399]
Explanations have various effects on users [. . .]. They can [. . .] increase the acceptance of recommendations [. . .].	[410]
The intuition is that if a user can query the system's decisions, s/he is less likely to abandon it and, indeed, may accept the system's choices over his/her own [. . .].	[412]
[. . .] explanations in a natural language are intuitive to humans, which can lead to a higher level of [. . .] willingness to use autonomous systems.	[422]
Explanations may increase user acceptance of the system or the given recommendations [. . .].	[465]
When they did occur evaluations of explanations have largely focused on user acceptance of the system [. . .] or acceptance of the systems' conclusions [. . .].	[465]

Continued on next page . . .

Table 6 – continued from previous page

Quote	Src.
Cramer et al. have investigated the effects of transparency on other evaluation criteria such as [...] acceptance of items [...].	[465]
Explanations have been found to increase user [...] acceptance of [...] the overall recommender system [...].	[465]
Research shows that explanations [...] improve user acceptance of recommendations [...].	[478]
Studies show that item-based explanations improve users' acceptance of recommendations [...].	[478]
Explanations describe the decision made by a machine learning model in order to gain user acceptance [...].	[488]
We list several types and goals of transparency. [...] To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system.	[490]
[Explanation] helps increase ES users' confidence in the system's problem-solving competence and hence, the acceptability of the conclusions.	[501]
[...] explanations positively impact the users' [...] commitment to repeatedly use [the system] and recommend it to others.	[502]
The findings show that the explanation feature can significantly increase a recommender system's perceived usefulness and thus contributes to increase users' repeated usage intention and their commitment to recommend the service.	[502]
[...] both the explanation of and trust in ML play significant roles in affecting the user acceptance of ML in practical applications.	[510]

Table 7: Quotes for the desideratum *Accountability*.

Quote	Src.
There has been increased attention into [...] accountable [...] algorithms [...], with [...] DARPA's Explainable AI (XAI) initiative [...].	[1]
[...] explanation approaches might serve regulatory goals of rendering algorithmic decision-making more [...] accountable [...].	[72]
The potential for [...] explanation systems to provide justice-related information, fulfilling the policy goals of [...] accountability [...], has recently been noted [...]	[72]
Counterfactuals are often used to determine legal culpability [...].	[95]
[The] relevance [of counterfactuals] to a variety of AI applications has been known for some time, ranging [...] from fault diagnosis to the determination of liability [...].	[95]
Algorithmic transparency provides several benefits. First, it is essential [...] to hold entities in the decision-making chain accountable [...].	[139]
Obviously, intermediate oversight authorities are to retain full rights of transparency as far as the model and its proxies are concerned, otherwise all accountability is gone.	[287]
[...] the second option (of foregrounding interpretability) opens up possibilities for fully fledged accountability.	[287]
While there are many approaches to increasing accountability in AI systems, we shall focus on one in this report: explanation [...].	[156]
[...] this principle, which we synthesise as "explicability" [...] in the ethical sense of "accountability" [...].	[175]
The role of explanation has been examined to enforce accountability under the law [...].	[189]
[...] outcome explanation had mixed effects, [...] reducing algorithmic accountability.	[302]
[...] outcome explanation [...] increased perceived fairness: it [...] made them attribute less accountability to algorithms in distributive outcomes.	[302]
Transparency, which refers to the understandability of a specific model, can be a mechanism that facilitates accountability.	[303]
Not only does provenance allow us to provide a form of explanation, it is a critical piece in achieving accountability as well.	[311]
[...] the increasingly widespread applicability of [...] models necessitates the need for explanations to hold such models accountable.	[322]
[...] we argue that, if xAI is to produce methods that make algorithmic decision-making systems more trustworthy and accountable, the field's attention must shift to the development of interactive methods for post-hoc interpretability [...].	[341]
[...] being able to explain its decisions [...] allows the system to be held accountable.	[345]
Transparency and the ability to explain AI decision making are core requirements for important aspects such as [...] accountability of algorithms.	[349]

Continued on next page ...

Table 7 – continued from previous page

Quote	Src.
Finally, a purely functional understanding of a model would also impede legal accountability [...] for the decisions of the model.	[365]
Much of the literature on transparency also emphasizes the goal of governing a system through accountability [...].	[388]
[...] different functions that transparency is thought to serve ([...] accountability).	[388]
The owner is concerned with explainability questions about the capabilities of the application, e.g. [...] aspects of accountability, e.g. to what extent can application malfunction be attributed to the DNN component?	[391]
Transparency is normally also a precondition for accountability: i.e. the extent to which the responsibility for the actionable outcome can be attributed to legally (or morally) relevant agents [...].	[391]
Explanations may even be the first step toward remedy, a critical aspect of accountability.	[395]
Only human beings can be held morally accountable so it should be human beings that are in control over these decisions [given by explanation].	[397]
An intelligent robot that is explainable yields several important advantages. [...] Accountability. As systems become more mission-critical, society will increasingly want to know where the blame lies when things go wrong.	[426]
[...] beyond debugging and accountability, explanations [...] help a user to understand [...].	[425]
It is vital to know if an explanation is Post-Hoc Rationalisation or Introspective when used for [...] compliance and accountability.	[427]
To be trusted, a system has to demonstrate [...] that the process leading to the decision is transparent and accountable [...]. Explanations form a vital part of satisfying these requirements.	[427]
[...] some [explainability methods] can also be used to assess accountability of the underlying predictive model [...].	[442]
The explainee can steer the explanatory process to [...] assess accountability [...].	[443]
[...] there is renewed interest in understanding the decisions of these algorithms [through explanations] as a means to [...] promote accountability.	[452]
The issue of AI legal accountability has recently been broached [...]. Interpretability and explainability of the system would come to the forefront of AI requirements in such a circumstance.	[476]
Abdul had identified other goals for XAI, such as providing transparency for algorithmic accountability [...].	[485]
We list several types and goals of transparency. [...] To provide an expert (perhaps a regulator) the ability to audit a prediction or decision [...]. This [...] will facilitate assignment of accountability and legal liability	[490]

Table 8: Quotes for the desideratum *Accuracy*.

Quote	Src.
[...] the most interpretable models usually are less accurate.	[4]
[...] intrinsic interpretable models come at a cost of accuracy.	[4]
[...] the technical challenge of explainability involving the tradeoff between accuracy and interpretability [...].	[4]
Explanations also help users to evaluate the accuracy of the system's predictions.	[29]
How much recommendation accuracy would one need to sacrifice by making a recommender system both transparent and scrutable?	[51]
Later studies [...] showed that explanations significantly increase users' [...] ability to correctly assess whether a prediction is accurate.	[73]
[...] model-level explanation makes users more likely to correctly predict the model's success with new samples.	[73]
To increase [...] accuracy in their training by designers, there is a need to enable AI systems to provide [...] explanations [...].	[95]
This transparency [caused by explanations] however comes at the cost of some classification accuracy.	[118]
"[...] "how" explanations [... enable] users to verify that an algorithm has accurately [...] ""produc[ed] and certif[ied] knowledge""."	[123]
Future research in software analytics should explicitly address the trade-off between explainability and prediction accuracy.	[131]
The use of simple models improves explainability but requires a sacrifice for accuracy.	[131]

Continued on next page ...

Table 8 – continued from previous page

Quote	Src.
[...] studies [...] found a positive relationship between frequency of novice use of explanations and problem solving performance – [...] in terms of accuracy of the quality of the decisions made [...].	[137]
[...] explanation [...] can lead to greater accuracy in the ensuing decision making [...].	[137]
[...] studies [...] show that explanation [...] can lead to greater accuracy in the ensuing decision making [...].	[137]
[...] in machine learning, one either goes with the flow of increasing accuracy and thereby sacrifices explanatory power [...].	[287]
We focus on the comprehensibility of classification models, rather than on the trade-off between predictive accuracy and comprehensibility.	[179]
[...] a rule-based classification system [i.e., an interpretable model] was proposed to trade-off classification accuracy and interpretability [...]	[187]
[...] they often achieve high interpretability with minimal sacrifice in classification accuracy [...].	[187]
Several other visualizations focus on improving the accuracy of recommendations with both explanations and support for user control.	[203]
There is an inherent tension between ML performance (predictive accuracy) and explainability	[236]
Often the best-performing methods such as DL are the least transparent, and the ones providing a clear explanation [...] are less accurate [...].	[236]
[...] transparency in design [...] may foster a better understanding of [...] the extent to which [the system] is fair and accurate.	[274]
However, explainability requirements may conflict with other softgoals such as [...] precision [...].	[277]
A number of motivations one might have in seeking explicability were cited earlier, such as [...] assurance of accuracy [...].	[280]
Concerns about [...] reassuring oneself about the accuracy of one's program all center around the epistemological notion of justification.	[280]
[...] models that are easy to interpret by humans [...] might yield lower accuracy [...].	[303]
The experimental results [...] have shown that In2Rec can effectively improve the [...] recommendation accuracy [...].	[307]
[...] (Intrinsic) explainability involves using a simpler model to fit data which can negatively affect predictive accuracy [...].	[322]
More generally, interpretability could contribute to the design of more accurate [...] classifiers.	[322]
Indeed, there is a well-known trade-off between accuracy and explainability [...].	[372]
In this paper, we use visual explanation for improving the accuracy of VQA systems.	[389]
It has been previously noted that an inverse relationship often exists between machine learning algorithms' accuracy and their explainability [...].	[399]
The goal [...] is to allow developers of XAI agents to trade off the need for explainability against other factors such as [...] predictive accuracy [...] of the underlying ML systems.	[425]
The contribution of our work is a set of categories that enable [...] to trade off the need for explainability against other factors such as [...] predictive accuracy [...] of the underlying ML systems.	[427]
Cramer et al. found that transparency led to changes in user behavior that ultimately decreased recommendation accuracy [...].	[465]
Research shows that explanations help users make more accurate decisions [...].	[478]
Studies show that item-based explanations [...] help users make accurate decisions [...].	[478]
Explanation also shows the ability to correctly assess whether a prediction is accurate [...].	[508]

Table 9: Quotes for the desideratum *Autonomy*.

Quote	Src.
In particular, [explanations] were presented to determine the level of autonomy to grant to an agent [...]	[30]
[...] for AI to promote and not constrain human autonomy, our "decision about who should decide" must be informed by knowledge of how AI would act instead of us [...].	[175]
Human-Autonomy Teaming (HAT) is required, where humans interact with the AI systems, and for this humans need to understand why the AI system is suggesting something that the human would not do: this requires interaction [and interaction requires explainable AI].	[177]

Continued on next page ...

Table 9 – continued from previous page

Quote	Src.
Some of the benefits provided [by explanation facilities] are: [. . .] allowing the user to add his knowledge and inference skills to the complete decision process.	[222]
[. . .] failure to render the processing logic comprehensible to data subject's disrespects their agency [. . .].	[342]
The analysis of explanation [. . .] has ethical consequences when we connect it to the notion of informed consent, which can be defined as "an autonomous authorisation by a patient or subject"	[373]
Transparency can empower users to make informed choices about how they use an algorithmic decision-making system [. . .].	[388]
[. . .] for AI to promote and not constrain human autonomy, our 'decision about who should decide' must be informed by knowledge of how AI would act instead of us [. . .].	[397]
AI will constrain rather than promote human autonomy unless we have the "knowledge of how AI would act instead of us" [. . .].	[397]

Table 10: Quotes for the desideratum *Confidence*.

Quote	Src.
The studies show that transparency and trust are going hand in hand to increase the user's confidence in the system by understanding how its reasoning mechanism works [. . .].	[30]
[. . .] purposes motivating the need for interpretable AI models, such as [. . .] confidence [. . .].	[54]
[. . .] explanations can serve a multiplicity of aims, such as inspiring the user's confidence in the system [. . .].	[51]
Later studies [. . .] showed that explanations significantly increase users' confidence [. . .].	[73]
[. . .] medical students [. . .] were more confident about disagreeing with [a system] when the explanations did not account adequately for all of the aspects of the case.	[101]
Systems researchers have emphasized the importance of explanations as a means of [. . .] increasing confidence in systems' abilities [. . .].	[123]
[. . .] explanations allow users to make inferences about a system's abilities and underlying motives, which form the basis of confidence [. . .] in a system.	[123]
[. . .] transparency [. . .] is necessary to build confidence in the system.	[127]
[. . .] global explanations seem to render more confidence in understanding the model [. . .].	[150]
Studies have explored [. . .] benefits related to providing explanations [. . .] many focusing on the advantage of earning users' [. . .]confidence in the systems [. . .].	[165]
Explainability [. . .] can help build [. . .] confidence [. . .].	[167]
AI rationalization has a number of potential benefits over other explainability techniques: [...] humanlike communication [. . .] may afford [. . .] advantages such as higher degrees of [. . .] confidence [. . .].	[166]
[. . .] transparency of music recommendations increased participants' [. . .] confidence [. . .].	[173]
If doctors want to use a neural network to make a diagnosis, they need to be confident that there is a clear rationale for the NN to diagnose a cancer [. . .].	[177]
Understanding a very accurate classification model might give us more confidence that the model is really capturing the correct patterns in the target domain [. . .].	[179]
[. . .] explanations [. . .] promote objectives such as [. . .] confidence in decision making [. . .].	[180]
[. . .] the intention behind disclosing the reasoning process of the system could be to increase the user's confidence in making the right decision [. . .].	[180]
"Indeed, explanations can differ significantly; for instance, they may attempt to maximize the user's confidence throughout the shopping experience [. . .]."	[180]
[. . .] confidence [. . . was] strongly affected by the presence of justification-type explanations [. . .].	[199]
We believe that by providing transparency into the workings of the ACF process, we will build users' confidence in the system [. . .].	[222]
Some of the benefits provided [by explanation facilities] are: [. . .] User understanding of the reasoning behind a recommendation, so that he may decide how much confidence to place in that recommendation.	[222]
[. . .] explainable-AI calls for confidence [. . .].	[236]

Continued on next page . . .

Table 10 – continued from previous page

Quote	Src.
[...] explanations [...] help users understand its reasoning and instill confidence [...].	[274]
Increased transparency is also associated with [...] higher confidence in system recommendations [...].	[274]
[...] explanations could help enhance decision confidence for follow-up actions [...].	. [304]
[...] motivations for explanations in recommender systems: [...] trust, which increases users' confidence towards recommendations.	[348]
The question of what kinds of explanation a human can utilize implies the presence of a downstream task. [...] extrinsic tasks include goals such as [...] trust – given the input, output, explanation, and observations of the world, does the explanation increase the human user's confidence in the agent?	[350]
Explanation Purposes Identified in Primary Studies [...]: [...] Increase users' confidence in the system.	[358]
The aim of local explanations is to strengthen the confidence and trust of users [...].	[391]
[...] explanations in a natural language are intuitive to humans, which can lead to a higher level of [...] confidence [...].	[422]
This is the goal of increasing confidence in the advice or solution offered by the system by giving some kind of support [in form of explanations] for the conclusion suggested by the system.	[444]
They suggest three major explanation goals. [...] Finally, the goal of ratification is to increase the end user's confidence in the system's conclusion.	[444]
Possible aims for explanations: [...] Increase users' confidence in the system.	[464]
Explanatory criteria and their definitions: [...] Increase users' confidence in the system.	[465]
Explanatory aims: [...] Increase users' confidence in the system.	[466]
[...] a detailed, full explanations may be "excessive" to the users [...], which had a negative impact on user confidence and enjoyment [..].	[473]
He contrasts explanation-for-trust [...] with explanation-for-confidence (i.e., explanation to make the user feel comfortable in using the system, by providing information on its external communications).	[479]
[...] non-expert users will need an explanation that increases their confidence and trust [...].	[479]
[...] an explanation for an end-user is intended to increase the user's confidence in the system [...].	[493]
There are three major explanation goals, namely [...] ratification. [...] in the context of ratification, the goal [...] is to increase the user's confidence in the expert system.	[493]
[Explanation] helps increase ES users' confidence in the system's problem-solving competence [...].	[501]
Other studies [...] consistently showed that explanations significantly increase users' confidence and trust [...].	[508]
Other studies [...] consistently showed that explanations significantly increase users' confidence and trust [...].	[510]

Table 11: Quotes for the desideratum *Controllability*.

Quote	Src.
[...] people should [...] feel in control. [...] prior work has identified issues [...] when this is not the case. [...] To address these problems, machine learning algorithms need to be able to explain how they arrive at their decisions.	[1]
Explainability [...] can also help prevent things from going wrong. Indeed, understanding more about system behavior [...] helps to rapidly identify and correct errors in low criticality situations (debugging). Thus enabling an enhanced control.	[4]
Explanations are also useful for control purposes.	[30]
Prior work shows that explanations [...] can increase [...] perception of control [...].	[96]
[...] information transparency enable[s] users to control their interactions more actively [...].	[112]
[...] interface features that inform users about the overt personalization mechanism [...] can increase perceived control [...].	[112]
In recent years, several interactive visualizations have been elaborated to support both explanations and user control over recommendations [...]	[203]
Several other visualizations focus on improving the accuracy of recommendations with both explanations and support for user control.	[203]
The result [from adding explanations] will be filtering systems [...] which give greater control to the user.	[222]

Continued on next page ...

Table 11 – continued from previous page

Quote	Src.
[...] people look for explanations to improve their understanding [...] so that they can derive stable model that can be used for prediction and control.	[339]
Transparency is generally desired because algorithms that are poorly predictable or explainable are difficult to control, monitor and correct [...].	[342]
Traditionally, computer programmers have had “control [...] ” insofar as they can explain its design and function to a third party [...].	[342]
The proposed visualization helps to create [...] control of personalized filtering to alleviate the "filter bubble" problem [...].	[348]
[...] explanations can represent a starting point for better user control [...].	[358]
Transparency mechanisms can convey a sense of iterative control [...].	[388]
I argue that a principle of explicability is primarily for the maintaining of meaningful human control over algorithms.	[397]
The idea is that an explanation of an algorithm’s output will allow a human being to have meaningful control over the algorithm [...]	[397]
To date, many reasons have been suggested for making systems explainable [...]: [...] To justify its decisions so the human participant can decide to accept them (provide control) [...].	[399]
Explainable artificial intelligence presented [...] mentions [...] control [...].	[417]
Explanations can help to teach users to [...] control a robot to achieve [a] task.	[426]
Explanations should be part of a cycle, where the user understands what is going on in the system and exerts control over the type of recommendations made [...].	[465]
[users’] perception of system explainability. However, the improvement comes with a price of reducing the user perception of control [...].	[473]
[...] a possibly overwhelming amount of information caused the users to decrease the perception of controllability	[473]
The finding of controllability and explainability trade-off is surprising, but not an uncharted area in the field of HCI.	[473]
When the overwhelming amount of information was provided [...] it impaired the user perception on controllability.	[473]

Table 12: Quotes for the desideratum *Debuggability*.

Quote	Src.
Explainability [...] can also help prevent things from going wrong. Indeed, understanding more about system behavior [...] helps to rapidly identify and correct errors in low criticality situations (debugging).	[4]
"[...] explainable AI planning (XAIP) is mostly algorithm dependent and serve more as a debugging system for an expert user."	[4]
Education and debugging were identified as the motivations of explanations, [...] the latter is considered for notifying users about the defects in the system [...].	[30]
Explanations were initially discussed [...] to support developers for system debugging.	[29]
Other related work [...] proposed generating explanations as a debugging tool.	[73]
[The] relevance [of counterfactuals] to a variety of AI applications has been known for some time, ranging [...] from fault diagnosis to the determination of liability [...].	[95]
Interpretability enables MLs models to be [...] debugged [...].	[99]
Interpretability also enables detection of faulty model behavior, through debugging [...].	[99]
Explanation can also play an important role in refining and debugging probabilistic systems.	[101]
These explanation generation techniques served more as a debugging system for an expert user [...].	[103]
Explainability is [...] potentially a trigger for new insights in the minds of practitioners. For example, a developer would want to understand why a defect prediction model suggests that a particular source file is defective so that they can fix the defect.	[131]
[...] trace explanations [...] are more likely to be used for debugging [...].	[137]
Explanation facilities can be of benefit in such systems because they can assist the user in either detecting or estimating the likelihood of errors in the recommendation.	[137]
[...] transparency can help detect errors in input data which resulted in an adverse decision [...]. Such errors can then be corrected.	[139]

Continued on next page ...

Table 12 – continued from previous page

Quote	Src.
In order to debug the procedure, he asks why the procedure led to the error, and not to the desired result (P-contrast question).	[497]
Explainable AI (XAI) is a field broadly concerned with making AI systems more transparent so people can [...] accurately troubleshoot it [...].	[150]
[...] using global explanations to understand and evaluate the model, and local explanations to scrutinize individual cases.	[150]
A key issue is whether explanations of system reasoning make it easier to detect erroneous recommendations—by, for instance, letting users discover flaws in the system's reasoning [...].	[165]
[...] it will be important to study whether and how explanations support users in reliably identifying system inaccuracies [...].	[165]
Employing such an abstraction [i.e., a form of explanation] showed potential for aiding program analysis [...].	[168]
Questions that experts in artificial intelligence (AI) ask opaque systems provide inside explanations, focused on debugging [...].	[189]
Explanations of visual systems could also aid in understanding network mistakes and provide feedback to improve classifiers.	[219]
The explanations can be useful [...] to allow him to understand how to modify its parameters if it does not behave as expected.	[221]
"[...] significant insights about an agent's behavior can be gained, e.g. explaining strange behavior, identifying ""bugs"" in preferences [...]."	[257]
"We provided case studies [...] to illustrate the potential of our visual explanations [...]. This allowed for spotting certain ""bugs"" [...]."	[257]
Explanations are also used in diagnosis. One might ask why a system failed and then repair a part to bring it back to its normal function.	[267]
These anecdotes [about an explanation tool] suggest that it may be possible to train developers to be more objective and careful about their debugging efforts by using the tool.	[276]
Explanations enable understanding and thereby [...] help in locating sources of error [...].	[277]
Part of enabling end users to "debug" their intelligent agents is explaining these agents to users well enough for them to build useful mental models.	[282]
To directly support end-user "debugging" of assistant behaviors [...], we present a Why-oriented approach which allows users to ask questions about how the assistantmade its predictions [and] provides answers to these "why" questions [...].	[283]
Our Why-oriented approach allows end users to debug such assistants by manipulating their underlying logic, not just their predictions [...].	[283]
Several informants [...] considered explanations as an integral part [...] to improve AI performance. Such needs are not only seen in debugging tools [...].	[304]
[...] generating explanations of application behavior [...] has been employed in [...] end-user debugging [...].	[305]
Our method generates contrastive explanations [...] and aims to help users understand (1) what contributed to the large error, and (2) what would be needed to produce a prediction with an acceptable error.	[312]
Our work also showcases the fact that interpretability is [...] a powerful tool for detecting flaws in the model [...].	[322]
The benefits of our proposed explainability framework are three-fold. We can [...] fix or delete adversarial examples [...].	[322]
[...] undesirable inferences can be subsequently debugged by users through machine coaching [i.e., a type of explainability].	[337]
Transparency is generally desired because algorithms that are poorly predictable or explainable are difficult to control, monitor and correct [...].	[342]
Explanations can be necessary to [...] verify and improve the functionality of a system (i.e. as a type of 'debugging' [...]) [...].	[341]
[...] explanations are vital in safety-critical systems, [...] to determine errors and faults in the system and to ensure that problems are fixed.	[345]
[...] such explanations were [...] being used in many cases only to support system debugging.	[358]
Explanation Purposes Identified in Primary Studies [...]: Debugging. [...] Allows users to identify that there are defects in the system.	[358]
[...] engineers are interested in explanations [...] that can be used for model debugging.	[391]
The information is used to [...] debug or repair the functioning of a system.	[391]
[...] explainability of this type is more appropriate for system debugging than for other uses.	[399]
To date, many reasons have been suggested for making systems explainable [...]: To explain the agent's choices to better [...] debug the system in previously unconsidered situations.	[399]

Continued on next page ...

Table 12 – continued from previous page

Quote	Src.
For debugging purposes, such developers must dig through the accumulated robot logs to find out about the robot experience in great detail.	[400]
An intelligent robot that is explainable yields several important advantages. [...] Ease of Debugging.	[426]
By making the decision processes transparent, it becomes easier for developers to discover and fix failures.	[426]
[...] the agent should be able to [...] give the user enough information to correct the error and understand its scope.	[426]
[...] beyond debugging and accountability, explanations [...] help a user to understand [...].	[425]
It is vital to know if an explanation is Post-Hoc Rationalisation or Introspective when used for correcting faults, predicting behaviour in critical systems [...].	[427]
Approaches that seek to explain such systems [...] are helpful to at least detect and identify the problem [of adversarial manipulation].	[427]
Furthermore, our explainability approach can help to identify errors [...] in the underlying machine learning model.	[441]
[...] some [explainability methods] can also be used to assess accountability of the underlying predictive model, e.g., debug and diagnose it [...].	[442]
The explainee can steer the explanatory process to [...] debug predictive models [...].	[443]
[...] goals in the ML domain, namely [...] diagnosis and refinement, [... are] related to the problems of interpretability of the ML results and comprehensibility of the obtained models.	[476]
If the system behaved unexpectedly or erroneously, users would want explanations for [...] debugging to be able to identify the [...] fault and [...] make corrections.	[485]
Explanations describe the decision made by a machine learning model [... for] debugging the machine learning system to identify flaws and inadequacies or distributional drift [...].	[488]
We list several types and goals of transparency. [...] For a developer, to understand how their system is working, aiming to debug or improve it [...].	[490]
Essentially, we want the robot to have "conversation" with a human so failures can be prevented [and] errors can be resolved [...]. There are two things that a human user can do here: Fully debugging the learned robot's plan [...].	[492]
ModelTracker [...] provides an intuitive visualization interface for [...] debugging.	[510]

Table 13: Quotes for the desideratum *Education*.

Quote	Src.
Education and debugging were identified as the motivations of explanations, the former is referred to as allowing users to learn something from the system [...]	[30]
[Explanations] can guide the user during the use of the system, working as a tutorial, to introduce the software features.	[108]
Explanations may also help to improve the usability of the system, [...] teaching the user how to better operate it.	[108]
The goal of explanation in such system should thus be to educate users about product knowledge by explaining what products do exist [...].	[110]
[...] the explanation for such system should be able to educate users about product knowledge [...]	[111]
If explanations can increase transparency and interpretability, this might improve [...] the pedagogical effectiveness [...].	[386]
[...] users will nonetheless benefit if some relevant details are made transparent because they will better be able to tell the uses and roles in inquiry for which the device is suited.	[130]
Learning is inhibited by a lack of time [...] – something that could be improved by easy access to explanations.	[137]
[...] by explaining why an adverse decision was made, it can provide guidance on how to reverse it [...].	[139]
Explanations, when suitably designed, have been shown to improve [...] learning.	[199]
Use of explanations aids learning (transfer of knowledge to non-KBS contexts).	[199]
Explanation use has been shown to have positive outcomes – [...] in some cases, improved learning.	[199]
Some of the benefits provided [by explanation facilities] are: [...] Education. Education of the user as to the processes used in generating a recommendation [...].	[222]
Educates Consumer: The process of providing good training explanations will help properly set expectations for what kind of explanations the system can realistically provide.	[225]

Continued on next page ...

Table 13 – continued from previous page

Quote	Src.
[...] self-explanation improves learning [...].	[231]
This suggests in situ explanations are a necessary condition to help end users learn how a machine learning system operates.	[281]
Our most complete explanations were associated with the highest perceived benefits and lowest perceived costs of learning about the system.	[282]
Engaging in explanation can facilitate learning [...].	[308]
And during learning, explanations are used to justify whether the newly solved cases should be learned, to determine which part of the case should be learned [...].	[325]
It is clear that the primary function of explanation is to facilitate learning [...].	[339]
[...] explanations have several functions other than the transfer of knowledge, such as [...] learning [...].	[339]
[...] explanation is good for learning and generalisation.	[339]
Explanations can be necessary to [...] help developers and humans working with a system learn from it [...].	[341]
[...] being able to explain its decisions allows the humans to learn [...].	[345]
[...] some benefits provided by explaining recommendations such as: [...] education [...].	[348]
Explanations are given to clarify, change or impart knowledge [...].	[357]
Explanation Purposes Identified in Primary Studies [...]: Education. [...] Allow users to learn something from the system.	[358]
Transparency mechanisms also function to help users to learn about how the system works [...].	[388]
Explanations can help to teach users to perform a similar task or control a robot to achieve this task.	[426]
[...] it is often beneficial for learning if the user participates in the formation of explanations [...].	[444]
They suggest three major explanation goals. [...] The system should [...] help the user learn the methods and knowledge used in the problem solving process.	[444]
After all, a true explanation teaches us something, or, if you like, we learn from it.	[471]
[...] non-expert users will need an explanation [...] that also teaches them how to use the system correctly and securely [...].	[479]
Essentially, we want the robot to have "conversation" with a human so [...] learning time can be reduced.	[492]
There are three major explanation goals, [...] duplication [...]. [...] In the context of duplication, the goal [...] is [...] to transfer [the expert system's] knowledge to the user.	[493]

Table 14: Quotes for the desideratum *Effectiveness*.

Quote	Src.
Explainable recommendation methods [...] improve effectiveness [...].	[43]
[...] explanations can serve a multiplicity of aims, such as [...] helping users make good decisions (effectiveness) [...].	[51]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: [...] effectiveness [...].	[72]
[...] a machine-learning based information system that can explain itself may allow more efficient and effective use of the technology.	[82]
[...] explanatory aim (e.g., [...] users' decision effectiveness).	[110]
We also performed an online user study [...], which shows our [explanation interface] can significantly increase users' decision effectiveness [...].	[111]
Users' objective decision effectiveness is also increased [by explanation interfaces] because almost half of them can make better choices [...].	[111]
[...] explanations are shown useful to improve users' decision effectiveness [...].	[111]
explanation purpose (e.g., [...] users' decision effectiveness)[...].	[111]
[...] explanations may be useful for increasing an intelligent system's overall effectiveness [...].	[386]
Critically, explanations [...] provide a more effective interface for the human in-the-loop [...].	[150]
[...] improved decision efficacy is an expected benefit for at least some explanations [...].	[165]
[...] the possible objectives of explanations are manifold, including aims such as increasing [...] effectiveness [...].	[180]
[...] effectiveness [... was] strongly affected by the presence of justification-type explanations [...].	[199]
Poorly designed explanations can actually decrease the effectiveness of a recommender system.	[222]

Continued on next page ...

Table 14 – continued from previous page

Quote	Src.
The result [from adding explanations] will be filtering systems that are [...] more effective [...].	[222]
[...] motivations for explanations in recommender systems: [...] effectiveness [...].	[348]
Explanation Purposes Identified in Primary Studies [...]: Effectiveness. [...] Help users make good decisions.	[358]
[...] explanations can help users make better decisions [...].	[358]
The second most frequent purpose of explanations [...] is effectiveness [...].	[358]
"[...] an organization-based explanation interface is likely to be more effective than the simple ""why"" interface [...]"	[382]
Initial evidence suggests that explanations may be useful for increasing an intelligent systems overall effectiveness [...].	[385]
Robots can collaborate more effectively with humans if they can describe their decisions, the underlying beliefs, and the experiences that informed these beliefs.	[452]
[...] possible aims when explaining the outcomes of an algorithm to users: [...] effectiveness[...].	[229]
Explanations can have many advantages, [...] helping users make good decisions.	[463]
Among other things, good explanations could [...] make it [...] easier for users to find what they want [...].	[463]
Among other things, good explanations could [...] make it [...] easier for users to find what they want [...].	[464]
Possible aims for explanations: Effectiveness. Help users make good decisions.	[464]
In this way, we distinguish between different explanation such as e.g. [...] explaining why the user may or may not want to try an item (effectiveness).	[465]
Explanatory criteria and their definitions: Effectiveness [...]. Help users make good decisions.	[465]
[...] good explanations may help users make better decisions.	[465]
Rather than simply persuading users to try or buy an item, an explanation may also assist users to make better decisions.	[465]
Explanatory aims: Effectiveness. Help users make good decisions.	[466]
[...] explanations can be more focused on helping users make decisions (about the items) that they are happy with: effectiveness.	[466]
Our design of tagsplanations is motivated by three goals: justification, effectiveness, and mood compatibility.	[478]
Explanations provide many benefits, from improving user satisfaction to helping users make better decisions.	[478]
[...] the effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users [...].	[485]
To enable end users to [...] effectively manage their intelligent partners, [...] researchers have produced many [...] algorithm visualizations, interfaces and toolkits [...].	[485]
We contend explaining why a recommended article is relevant will increase efficacy [...] in the generation of literature reviews [...].	[489]
However, in order for humans to [...] effectively manage the emerging AI systems, an AI needs to be able to explain its decisions and conclusions.	[505]
Thus, it is important to build more explainable AI, so that humans can [...] effectively manage the emerging AI systems [...].	[505]
Explainable ML aims to [...] enable human users to understand, appropriately trust, and effectively manage the ML-based solutions [...].	[508]
[...] both interaction with and transparency of the system help humans make effective uses of the AI system for trusting decisions.	[510]

Table 15: Quotes for the desideratum *Efficiency*.

Quote	Src.
In fact, requiring every AI system to explain every decision could result in less efficient systems [...].	[4]
For collaborative tasks, explanations were deemed essential to increase efficiency [...].	[30]
[...] explanations can serve a multiplicity of aims, such as [...] helping users [...] make decisions faster (efficiency) [...].	[51]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: [...] efficiency [...].	[72]
[...] a machine-learning based information system that can explain itself may allow more efficient and effective use of the technology.	[82]

Continued on next page ...

Table 15 – continued from previous page

Quote	Src.
[Explanations] may also help users to understand better all available features and support in completing tasks faster.	[108]
[...] explanatory aim (e.g., [...] users' decision efficiency [...]).	[110]
As for decision efficiency, [the study] shows users spent more time in making decisions [with explanation interfaces].	[111]
[...] explanation purpose (e.g., [...] users' decision efficiency [...])[...].	[111]
Although the interactive approach is more effective at improving comprehension, it comes with a trade-off of taking more time.	[113]
[...] studies [...] found a positive relationship between frequency of novice use of explanations and problem solving performance – [...] in terms of [...] time taken to make decisions.	[137]
[...] explanation [...] can [...] reduce time in decision making.	[137]
[...] studies [...] show that explanation [...] can [...] reduce time in decision making.	[137]
[...] user performance (i.e., time to make a decision [...]) [... are] likely benefits of explanation facilities for recommender systems.	[137]
Secondly, full transparency concerning the machine learning models in use may invite those concerned to game the system and thereby undermine its efficiency.	[287]
[...] explainable interfaces also have a positive effect on time [...].	[151]
However, society cannot demand an explanation for every decision, because explanations are not free. Generating them takes time and effort	[156]
The proposed concept performed significantly better compared to [...] simple explanations in terms of our [...] side goals to increasing perceived efficiency [...].	[291]
More generally, interpretability could contribute to the design of more [...] efficient classifiers	[322]
[...] motivations for explanations in recommender systems: [...] efficiency [...].	[348]
Looking at the historical developments, we can observe [...] that other potential purposes [of explanations], like [...] efficiency, received more attention in the recent past.	[358]
Explanation Purposes Identified in Primary Studies [...]: Efficiency. [...] Help users make decisions faster.	[358]
[...] explanations increase the user's ability to predict the classifier decision, while decreasing the time needed to reach a judgement.	[366]
[...] participants felt that [an organization-based explanation interface] would be easier for them to [...] make a quicker decision.	[382]
The goal [...] is to allow developers of XAI agents to trade off the need for explainability against other factors such as efficiency [...] of the underlying ML systems.	[425]
The contribution of our work is a set of categories that enable [...] to trade off the need for explainability against other factors such as efficiency [...] of the underlying ML systems.	[427]
[...] possible aims when explaining the outcomes of an algorithm to users: [...] efficiency [...].	[229]
Among other things, good explanations could [...] make it quicker [...] for users to find what they want [...].	[463]
Among other things, good explanations could [...] make it quicker [...] for users to find what they want [...].	[464]
Possible aims for explanations: Efficiency. Help users make decisions faster.	[464]
Explanations can also serve other aims such as [...] make it quicker [...] for users to find what they want [...].	[465]
Explanatory criteria and their definitions: Efficiency [...]. Help users make decisions faster.	[465]
Explanations may make it faster for users to decide which recommended item is best for them.	[465]
Explanatory aims: Efficiency. Help users make decisions faster.	[466]
We contend explaining why a recommended article is relevant will increase [...] efficiency in the generation of literature reviews [...].	[489]

Table 16: Quotes for the desideratum *Fairness*.

Quote	Src.
There has been increased attention into [...] fair [...] algorithms [...], with [...] DARPA's Explainable AI (XAI) initiative [...].	[1]

Continued on next page ...

Table 16 – continued from previous page

Quote	Src.
Using XAI systems [. . .] ensures that there is [. . . a] way to defend algorithmic decisions as being fair [. . .].	[4]
It starts with with the initial set of available discriminatory paths and generates other inputs belonging to nearby execution paths, thereby systematically performing local explainability while banking on the adversial robustness property.	[6]
Interpretability helps ensure impartiality in decision-making, i.e. to detect, and consequently, correct from bias in the training dataset.	[54]
[. . .] explainability can be considered as the capacity to reach and guarantee fairness in ML models.	[54]
[. . .] purposes motivating the need for interpretable AI models, such as [. . .] fairness [. . .].	[54]
[. . .] an explainable ML model suggests a clear visualization of the relations affecting a result, allowing for a fairness [. . .] analysis of the model at hand [. . .].	[54]
The literature also exploses [sic!] that XAI proposals can be used for bias detection.	[54]
[. . .] explanation approaches might serve regulatory goals of rendering algorithmic decision-making more fair [. . .].	[72]
The potential for [. . .] explanation systems to provide justice-related information, fulfilling the policy goals of [. . .] fairness, has recently been noted [. . .].	[72]
Requiring organisations to explain the logic behind their algorithmic decision-making systems [. . .] enables affected individuals to assess whether the logic of the system is just [. . .], which in turn might moderate their assessments of fairness of the decision outcomes [. . .].	[72]
Finding ways to reveal something of the internal logic of an algorithm can address concerns about lack of ‘fairness’ and discriminatory effects, sometimes with reassuring evidence of the algorithm’s objectivity.	[92]
Careful constraints on counterfactuals are required to provide interpretable models of the decisions of AI systems that people can [. . .] consider to be fair [. . .].	[95]
[. . .] explaining a ML model’s decisions provides a way to check [. . .] fairness [. . .].	[99]
[. . .] explanations are of uttermost importance to ensure algorithmic fairness, to identify potential bias/problems in the training data [. . .].	[99]
[. . .] desiderata that can be optimized through interpretability: [. . .] Fairness — Ensure that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups.	[99]
"[. . .] “how” explanations [. . . enable] users to verify that an algorithm has [. . .] fairly ""produc[ed] and certif[ied] knowledge""."	[123]
Algorithmic transparency provides several benefits. First, it is essential to enable identification of harms, such as discrimination, introduced by algorithmic decision-making [. . .].	[139]
Critically, explanations [. . .] provide a more effective interface for the human in-the-loop, enabling people to identify and address fairness and other issues.	[150]
Our results highlight the need to provide different styles of explanation tailored for exposing different fairness issues.	[150]
[. . .] we show that local explanations are more effective in exposing fairness discrepancies between different cases, while global explanations seem to [. . .] generally enhance the fairness perception.	[150]
Explainable AI (XAI) is a field broadly concerned with making AI systems more transparent so people can [. . .] accurately troubleshoot it, fairness issues included.	[150]
Interpretability is used to confirm other important desiderata of ML systems: [. . .] Notions of fairness or unbiasedness imply that protected groups (explicit or implicit) are not somehow discriminated against.	[155]
[. . .] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
Develop auditing mechanisms for AI systems to identify unwanted consequences, such as unfair bias [. . .].	[175]
These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data [. . .].	[188]
May Reduce Bias: Providing explanations will increase the likelihood of detecting bias in the training data [. . .].	[225]
Uncovering the bias of the ranking algorithm [via explantions] will help researchers to better support their research results.	[227]
Without good models and the right tools to interpret them, data scientists risk making decisions based on hidden biases, spurious correlations, and false generalizations.	[233]
Lacking an explanation for how models perform can lead to biased and ill-informed decisions [. . .].	[233]
[. . .] explainable-AI calls for [. . .] fairness [. . .].	[236]
We explained how to measure discrimination in data or decisions output by a classifier by explicitly considering explainable and illegal discrimination.	[259]

Continued on next page . . .

Table 16 – continued from previous page

Quote	Src.
[...] we presented the local techniques that remove exactly the illegal discrimination, allowing the differences in decisions to be present as long as they are explainable.	[259]
[...] transparency in design [...] may foster a better understanding of [...] the extent to which [the system] is fair and accurate.	[274]
Many reasons have been given to explain why "explicability," "interpretability," and/or "transparency" are important desiderata: [...] If we do not know how ML algorithms work, we cannot check or regulate them to ensure that they do not encode discrimination against minorities [...].	[280]
A number of motivations one might have in seeking explicability were cited earlier, such as [...] non-discrimination.	[280]
[...] outcome explanation had mixed effects, increasing or decreasing perceived fairness [...].	[302]
[...] outcome explanation [...] increased perceived fairness: it allowed people to understand equalities in utility distribution and the role of individual input [...].	[302]
[...] outcome explanation had mixed effects, increasing or decreasing perceived fairness [...].	[302]
[...] transparency decreased perceived fairness: outcome explanation made participants recognize uneven distributions and revealed differences in strategies across participants.	[302]
[...] users also deem explanations of the AI's decision as potential mitigation of their own decision biases.	[304]
Our work also showcases the fact that interpretability is [...] a powerful tool for detecting [...] biases in data [...].	[322]
Explainability in general also helps to identify bias in raw data [...].	[322]
Such debugging [based on explanation] will [...] not fully eliminate bias, but will, instead, replace it with bias that aligns better with each user's own preferences and beliefs.	[337]
[...] auditing can [...] detect discrimination or similar harms.	[342]
[...] the suggestions [...] must be perceived to be fair [...], and explanations are key to this.	[358]
Transparency mechanisms also enable users to identify biases that may result in negative consequences [...].	[388]
The aim of local explanations is to strengthen the confidence and trust of users that the system is not (or will not be) conflicting with their values, i.e. that it does not violate fairness or neutrality.	[391]
It is unfair that we can receive a low credit score, end up on a police watch list, get higher prison sentences, etc. without explanation about the considerations that led to those decisions.	[397]
To date, many reasons have been suggested for making systems explainable [...]: To explain the agent's choices to ensure fair [...] decisions are made [...].	[399]
[...] explanations include those designed for legal and policy experts to confirm that the decisions/actions of agent fulfill legal requirements such as being fair [...].	[399]
[...] reasons why ML interpretability is desired, namely [...] fair [...] decision making.	[417]
Sokol, Kacper, and Peter A. Flach - Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements (2018)	[441]
[...] some [explainability methods] can also be used to assess accountability of the underlying predictive model, e.g., [...] demonstrate its fairness [...].	[442]
The explainee can steer the explanatory process to inspect fairness (e.g., identify biases towards protected groups) [...].	[443]
[...] this meta-analysis shows that explanations do affect perceptions of fairness [...].	[472]
Exploratory fairness analysts might manually examine mechanisms behind a model's core logics and ask if they made sense.	[475]
The question is where does discrimination occur? The answer lies in the explainability [...].	[481]
[...] we developed five explanation strategies to mitigate decision biases [...].	[485]
Abdul had identified other goals for XAI, such as providing transparency for [...] detecting model bias [...].	[485]
We selected a subset of heuristic biases for which we identify how XAI can play a role to mitigate them [...].	[485]
Legally, opacity prevents regulatory bodies from determining whether a particular system processes data fairly and securely [...]. [...] Investigators within the Explainable Artificial Intelligence [...] research program intend to ward off these consequences [...].	[503]
Often knowing the reasons why a particular decision has been taken [through explanations ...] can engender [...] confidence that the people in charge of the process acted fairly and reasonably [...].	[506]

Table 17: Quotes for the desideratum *Informed Consent*.

Quote	Src.
The importance given to the information requirement, associated with transparency in the GDPR, reflects underlying assumptions about the value of informed consent for technology users.	[173]
Meaningful consent to data-processing is not possible when opacity precludes risk assessment [...].	[342]
The analysis of explanation [...] has ethical consequences when we connect it to the notion of informed consent [...].	[373]
Only if the right kind of information is given can informed consent on using the system and its outputs be established [...].	[373]
The security of a system thus needs to be explained to the user in order to allow her to make an informed decision on whether to use it.	[373]

Table 18: Quotes for the desideratum *Legal Compliance*.

Quote	Src.
Recently, the European Union approved a data protection law [...] that includes a "right to explanation" [...]	[1]
[...] AI needs to provide justifications in order to be in compliance with legislation [...].	[4]
[...] the "right to explanation", which is a regulation included in the General Data Protection Regulation (GDPR) [...].	[4]
[...] the [...] recent General Data Protection Regulation (GDPR) law [...] underlines the right to explanations [...].	[30]
[...] the European Union's new General Data Protection Regulation (GDPR), which also enforces a "right to explanation" (regarding specific algorithmic decisions) [...].	[43]
[Explainable Artificial Intelligence] is part of a context where laws reinforce the right of users [...].	[48]
[...] explanation approaches might serve regulatory goals [...].	[72]
In so far as these notions of justice capture the aims of the regulatory requirements, they may provide appropriate ways of measuring the adequacy of different explanation systems for these purposes.	[72]
Indeed, in many cases, [the ability to generate explanations] may even be required by law [...].	[103]
[...] we seek to learn design principles for explanation interfaces that communicate how decision-making algorithms work, in order to [...] support users' "right to explanation".	[113]
[...] explanations may be useful for [...] increased compliance [...].	[386]
[Having Transparency] may play a significant role in litigation.	[127]
For example, the EU General Data Protection Regulation (GDPR) requires organizations deploying ML systems to provide affected individuals with meaningful information about the logic behind their outputs.	[150]
[...] it is imperative – for [...] legal reasons [...] – that intelligent decision support systems provide users with access to the underlying models [...].	[165]
The importance given to the information requirement, associated with transparency in the GDPR, reflects underlying assumptions about the value of informed consent for technology users.	[173]
Within data protection law, notice and consent refers to providing information about the envisaged data processing to an individual [...].	[173]
There are growing legal implications in the use of AI, and in the cases where the AI system makes the wrong decision, or simply disagrees with the human, it is important to understand why a wrong or different decision was made: this is transparency.	[177]
[...] in some application domains users need to understand the system's recommendations enough to legally explain the reason for their decisions to other people.	[179]
These outside explanations can be used to [...] comply with regulatory and policy changes [...].	[189]
Similar recommendations in using explanations in law have been examined in promoting [...] liability for machines [...].	[189]
[...] the European Union's General Data Protection Regulation (GDPR) creates obligations for automatic decision making processes [...], with a provision including right to explanation.	[189]
When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them.	[193]
The GDPR's policy on the right of citizens to receive an explanation for algorithmic decisions highlights the pressing importance of human interpretability in algorithm design.	[193]

Continued on next page ...

Table 18 – continued from previous page

Quote	Src.
Explanation technologies are an immense help to companies for creating [. . .] products, and better managing any possible liability they may have.	[201]
[. . .] there is a growing demand that these systems provide explanations for their decisions, so that [. . .] a citizen's due process rights are respected [. . .].	[225]
"Indeed, a proposed regulation before the European Union [. . .] asserts that users have the ""right to an explanation""."	[231]
Articles 13 and 22 state a "right to explanation" for any algorithm whose decision impacts a person's legal status [. . .].	[233]
AI and policy scholars expect explanations to be important in future regulations of AI systems [. . .].	[233]
Moreover, transparency can be viewed as a regulatory or voluntary requirement.	[241]
Transparency might be seen as a regulatory requirement because laws and regulations may require organisations to be transparent for certain reasons and on certain processes [. . .].	[241]
What mattered [. . .] was to make sure that the information needed for rational self-governance was not concealed or unjustly controlled [. . .] – hence the recurrent campaigns for transparency and the right to know in many democratic societies.	[251]
[. . .] it is not obvious what policymakers or other stakeholders actually mean when they demand explainability and enshrine it in laws or guidelines.	[277]
Many reasons have been given to explain why "explicability," "interpretability," and/or "transparency" are important desiderata: [. . .] If we do not know how ML algorithms work, we cannot check or regulate them to ensure that they do not encode discrimination against minorities [. . .].	[277]
People have a right to know why an ML algorithm has produced some verdict (such as lack of creditworthiness) about them [. . .].	[280]
More recently, European privacy law codified the "right to an explanation" for users of platforms [. . .].	[302]
[. . .] secrecy [as the opposite of explainability][. . .] prevents violations of legal restrictions on disclosure of data.	[303]
The European Union introduced a right-to-explanation in GDPR as an attempt to remedy the potential problems given the rising importance of ML algorithms.	[322]
These are not mundane regulatory tasks, the provisions highlighted above can be interpreted to mean automated decisions must be explainable to data subjects.	[342]
Explanations can be necessary to comply with relevant legislation [. . .].	[341]
[. . .] explanations are vital in safety-critical systems, particularly where explanations are necessary for compliance [. . .].	[345]
Transparency and the ability to explain AI decision making are core requirements for important aspects such as [. . .] liability [. . .] of algorithms.	[349]
Finally, a purely functional understanding of a model would also impede legal accountability [. . .] for the decisions of the model.	[365]
Currently there is much debate regarding the safety of and trust in data processes in general, leading to investigations regarding the explainability of AI-supported decision making. The level of concern about these topics is reflected by offcial regulations such as the General Data Protection Regulation (GDPR)	[391]
[. . .] the offered explanations should match (within certain limits) the particular user's capacity for understanding [. . .], as indicated by the GDPR.	[391]
[. . .] the system's explanation is [. . .] to confirm that a secondary, legal, requirement is being met.	[394]
Furthermore, these explanations might be necessary for legal considerations [. . .].	[399]
To date, many reasons have been suggested for making systems explainable [. . .]: To explain the agent's choices to ensure [. . .] legal decisions are made [. . .].	[399]
[. . .] explanations include those designed for legal and policy experts to confirm that the decisions/actions of agent fulfill legal requirements [. . .].	[399]
It is vital to know if an explanation is Post-Hoc Rationalisation or Introspective when used for [. . .] compliance and accountability.	[427]
In the new age of GDPR [. . .], automated systems are now legally required (in Europe) to be able to explain how recommendations were computed.	[436]
[. . .] decisions about humans without them knowing drawn attention of lawmakers and regulators leading to DARPA's Explainable AI (XAI) project and European Union's General Data Protection Regulation (GDPR) [. . .].	[441]

Continued on next page . . .

Table 18 – continued from previous page

Quote	Src.
This [...] will be of particular importance [. . .] for [. . .] explainability methods, especially if compliance with best practices or legal regulations is required, e.g., the “right to explanation” introduced by the European Union’s General Data Protection Regulation (GDPR) [. . .].	[442]
[. . .] interpretability might be seen as the way to make model performance and guidelines compliance compatible.	[476]
Enforced in May 2018, [the GDPR] mandates a right to explanation of all decisions made by automated or artificially intelligent algorithmic systems [. . .].	[476]
Explanations describe the decision made by a machine learning model [. . . for] legality purposes that come from [. . .] the right to be informed about the basis of the decision [. . .].	[488]
We list several types and goals of transparency. [. . .] To provide an expert (perhaps a regulator) the ability to audit a prediction or decision [. . .]. This [. . .] will facilitate assignment of accountability and legal liability	[490]
Legally, opacity [. . .] may hinder end users from exercising their rights under the European Union’s General Data Protection Regulation [. . .]. [. . .] Investigators within the Explainable Artificial Intelligence [. . .] research program intend to ward off these consequences [. . .].	[503]
"Transparency can thus [. . .] come to embody an end or democratic value in its own right, a ""right to know"" [. . .]."	[506]

Table 19: Quotes for the desideratum *Morality*.

Quote	Src.
Using XAI systems [. . .] ensures that there is [. . . a] way to defend algorithmic decisions as being [. . .] ethical.	[1]
[. . .] an explainable ML model suggests a clear visualization of the relations affecting a result, allowing for a[n . . .] ethical analysis of the model at hand [. . .].	[54]
Work on perceptions of justice reveals much about the role of explanations in ethical assessments of human decisions.	[72]
We argue that the need for interpretability stems from an incompleteness in the problem formalization [. . .]. Below are some illustrative scenarios: Ethics: The human may want to guard against certain kinds of discrimination, and their notion of fairness may be too abstract to be completely encoded into the system (e.g., one might desire a ‘fair’ classifier for loan approval).	[155]
[. . .] it is imperative – for ethical [. . .] reasons [. . .] – that intelligent decision support systems provide users with access to the underlying models [. . .].	[165]
[. . .] this principle, which we synthesise as “explicability” [. . .] in the ethical sense [. . .].	[175]
Similar recommendations in using explanations in law have been examined in promoting ethics for design [. . .].	[189]
The main purpose to provide explainability of a model also varies, e.g., ethical reasons [. . .].	[234]
[. . .] explainable-AI calls for [. . .] ethics [. . .].	[236]
Last but not least, informants reflected on their ethical responsibilities to provide explanation.	[304]
[. . .] XAI is a key part of applying ethics to AI [. . .].	[322]
[. . .] transparency is often naively treated as a panacea for ethical issues arising from new technologies.	[342]
The analysis of explanation [. . .] has ethical consequences [. . .].	[373]
This points to the ethical issue of ensuring that the outputs of algorithms are not made based upon ethically problematic or irrelevant considerations. We expect, for example, a rejection for a loan not to be based on the color of the applicant’s skin (or a proxy thereof). An explanation of the algorithm’s decision can allow for someone to accept, disregard, challenge, or overrule the rejection.	[397]
It should be clear that explicability is considered to be an important part of achieving so-called ‘ethical’ [. . .] AI.	[397]
Microsoft, Google, the World Economic Forum, the draft AI ethics guidelines for the EU commission, etc. all include a principle for AI that falls under the umbrella of ‘explicability’.	[397]
To date, many reasons have been suggested for making systems explainable [. . .]: To explain the agent’s choices to ensure [. . .] ethical [. . .] decisions are made [. . .].	[399]
[. . .] explanations include those designed for legal and policy experts to confirm that the decisions/actions of agent fulfill legal requirements such as being [. . .] ethical [. . .].	[399]
[. . .] reasons why ML interpretability is desired, namely [. . .] ethical decision making.	[417]

Continued on next page . . .

Table 19 – continued from previous page

Quote	Src.
Explanations describe the decision made by a machine learning model [... for] legality purposes that come from ethical standards [...].	[488]

Table 20: Quotes for the desideratum *Performance*.

Quote	Src.
For collaborative tasks, explanations were deemed essential to increase [...] team performance [...].	[30]
This is true in the sense that there is a trade-off between the performance of a model and its transparency [...].	[54]
Explainability will also enhance trust in the system at the level of the users, and due to that improve the quality and performance of human-machine systems.	[43]
An explanation of a decision intended to help the user understand the AI system and make inferences about its future performance could best rely on better-world counterfactuals.	[95]
Software engineers need to pay attention to how the explanations will be integrated, without undermining system performance.	[108]
Providing an explanation [...] results in [...] enhanced user performance when using a system [...].	[123]
[...] making adaptive systems more transparent to the user could lead to [...] increases in system performance [...].	[124]
The main benefits arising from the use of explanation facilities are [...] user performance.	[137]
[...] studies [...] found a positive relationship between frequency of novice use of explanations and problem solving performance [...].	[137]
[...] user performance [... are] likely benefits of explanation facilities for recommender systems.	[137]
[...] transparency also appears as occluding performativity [...].	[173]
Explanations, when suitably designed, have been shown to improve performance [...].	[199]
Use of explanations improves the performance achieved with a KBS as an aid.	[199]
[...] explanations of all types was related to improved problem solving performance.	[199]
Explanation use has been shown to have positive outcomes – better performance [...].	[199]
Explanation capabilities [...] may improve the filtering performance of people using ACF systems.	[222]
We believe that explanations can increase the filtering performance.	[222]
[...] explanations must present easy-to understand coherent stories in order to ensure [...] good performance of the human-machine work system.	[231]
There is an inherent tension between ML performance [...] and explainability.	[236]
The benefits we see for including explanations is more information and knowledge for the user, thus [...] improving [...] task performance.	[254]
Providing explanations can increase performance on information retrieval tasks [...].	[274]
However, explainability requirements may conflict with other softgoals such as performance [...].	[277]
A less explainable system [...] could offer a higher performance.	[277]
Several informants [...] considered explanations as an integral part [...] to improve AI performance.	[304]
[...] explainability will also enhance trust at the user side, and because of that improve the human-machine interaction performance.	[349]
[...] explanations have a positive effect for the ability to predict the system's performance correctly [...].	[366]
Explanations improve [...] overall performance [...].	[388]
Explanation can be helpful in [...] leveraging systems for improving performance.	[389]
[...] engineers are interested in explanations of a functional nature, e.g. the effects of various hyperparameters on the performance of the network [...].	[391]
[...] an explainable model might even be done a a sacrifice to the system's performance.	[394]
[...] such a requirement [of explicability would] trade off the power of AI in terms of performance [...].	[397]
However, explainability is a trade-off, often occurring at the expense of other factors such as performance or development effort.	[427]

Continued on next page ...

Table 20 – continued from previous page

Quote	Src.
We examine the results in an attempt to ascertain whether explanation can be correlated with improved performance and/or user satisfaction. Our earlier results demonstrated clearly that the use of argumentation-based dialogue improves both of these measures.	[436]
[...] sometimes the decrease in predictive performance that is associated with making a particular system more explainable may not be worth it.	[442]
All of the explainability approaches should be accompanied by a critical discussion of performance – explainability trade-offs that the user has to face.	[442]
[...] this meta-analysis shows that explanations do affect [...] test performance.	[472]
[...] interpretability might be seen as the way to make model performance and guidelines compliance compatible.	[476]
Often we constrain the model to derive explanations. [...] This decreases the predictive power [...].	[481]
[...] users who are recipients of system-generated recommendations [...] prefer information that indicate the systems prior performance [...].	[482]
[...] software developers may be unable to intervene in order to quickly and systematically improve performance [...]. [...] Investigators within the Explainable Artificial Intelligence [...] research program intend to ward off these consequences [...].	[503]
ModelTracker [...] provides an intuitive visualization interface for ML performance analysis [...].	[510]

Table 21: Quotes for the desideratum *Persuasiveness*.

Quote	Src.
In addition to improving user acceptance of recommendations (persuasiveness), explanations can serve a multiplicity of aims [...].	[51]
Arguably, the most important contribution of explanations is not to convince users to adopt recommendations [...].	[70]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: [...] persuasiveness [...].	[72]
[...] explanation can be [...] for increasing [...] persuasiveness [...].	[110]
[...] explanatory aim (e.g., [...] system persuasiveness [...]).	[110]
[...] explanation can be [...] for increasing [...] persuasiveness [...].	[111]
[...] explanation purpose (e.g., [...] system persuasiveness [...])[...]	[111]
Explanations could also help convince users of a system's competence, if an explanation itself is acceptable to them.	[124]
The perceptions arising from the use of explanation facilities include [...] persuasiveness about the quality of the expert system.	[137]
[...] persuasiveness [... are] likely benefits of explanation facilities for recommender systems.	[137]
[...] feedback explanation facilities can be of benefit to expert decision making – leading to greater adherence to the system recommendation.	[137]
[...] the degree to which a user would be convinced to buy [are] likely benefits of explanation facilities for recommender systems.	[137]
Numerous studies have demonstrated the benefits of explanations in intelligent systems [...], including [...] making recommendations more persuasive [...].	[165]
[...] studies [...] may focus on how explanations enhance perceived [...] credibility of advice [...].	[165]
A key issue is whether explanations [...] make both correct and erroneous recommendations appear more plausible.	[165]
[...] the possible objectives of explanations are manifold, including aims such as increasing [...] persuasiveness [...].	[180]
Explanations that conform to Toulmin's model should be more persuasive [...].	[199]
[...] explanations could also convince users to invest in the system [...].	[304]
[...] explanations have several functions other than the transfer of knowledge, such as persuasion [...].	[339]
With respect to explanation in AI, persuasion is surely of interest: if the goal of an explanation from an intelligent agent is to generate trust from a human observer, then persuasion that a decision is the correct one could in some case be considered more important than actually transferring the true cause.	[339]
[...] motivations for explanations in recommender systems: [...] persuasiveness, which convinces users to form target attitude [...].	[348]

Continued on next page ...

Table 21 – continued from previous page

Quote	Src.
[...] explanations can [...] persuade [users] to make one particular choice [...].	[358]
Explanation Purposes Identified in Primary Studies [...]: Persuasiveness. [...] Convince users to try or buy.	[358]
Persuasiveness [...] was also in the focus of a number of studies [examining the purposes of explanation].	[358]
In this commercial context, the potential persuasive nature of explanations [...] also attracted more research interest recent years.	[358]
Traum et al. explained the justification within choices of their training agent to better convince the trainee [...].	[394]
Rosenfeld and Kraus created agents that use argumentation to better persuade people to engage in positive behaviors [...].	[394]
[...] the information will likely need to persuade the person to choose a certain action.	[399]
[...] the agent might need to provide information about its decision to help convince the human participant of the correctness of their solution [...].	[399]
The expected impacts of [...] explanations are as follows: [...] persuasiveness: recognition of a suitable context for usage motivates users to consume items [...].	[410]
We further confirmed that the hybrids of the context style and other explanation styles improve persuasiveness and usefulness.	[410]
With respect to personalization, we add the goal of persuasion, which aims at changing someone's beliefs through reasoning and argument.	[417]
An intelligent robot that is explainable yields several important advantages. [...] the agent should be able to [...] convince the user that the agent is correct [...].	[426]
[...] possible aims when explaining the outcomes of an algorithm to users: [...] persuasiveness[...].	[229]
Among other things, good explanations could [...] persuade [users] to try or purchase a recommended item.	[463]
Among other things, good explanations could [...] persuade [users] to try or purchase a recommended item.	[464]
Possible aims for explanations: Persuasiveness. Convince users to try or buy.	[464]
Explanations can also serve other aims such as [...] persuade [users] to try or purchase a recommended item.	[465]
Explanatory criteria and their definitions: Persuasiveness [...]. Convince users to try or buy.	[465]
Rather than simply persuading users to try or buy an item, an explanation may also assist users to make better decisions.	[465]
Cramer et al. have investigated the effects of transparency on other evaluation criteria such as [...] persuasion (acceptance of items) [...].	[465]
Explanatory aims: Persuasiveness. Convince users to try or buy.	[466]
[...] the most important contribution of explanations is not to convince users to accept the customized results [...]	[487]
We list several types and goals of transparency. [...] To lead a user (the audience) into some action or behavior [...].	[490]

Table 22: Quotes for the desideratum *Privacy*.

Quote	Src.
[...] one of the byproducts enabled by explainability in ML models is its ability to assess privacy.	[54]
[...] explaining a ML model's decisions provides a way to check [...] privacy [...].	[99]
[...] desiderata that can be optimized through interpretability: [...] Privacy — Ensure that sensitive information in the data is protected.	[99]
[...] designers may [...] increase system transparency to [...] lower privacy concerns toward the application [...].	[112]
[...] transparency can both help and hurt privacy [...]	[127]
[...] transparency can both help and hurt privacy [...]	[127]
First, for the sake of privacy it would be unwise to make underlying datasets freely available to anyone	[287]
Interpretability is used to confirm other important desiderata of ML systems: [...] Privacy means the method protects sensitive information in the data.	[155]
[...] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
Similar recommendations in using explanations in law have been examined [...] for privacy [...].	[189]

Continued on next page ...

Table 22 – continued from previous page

Quote	Src.
[...] a few participants said that interpretability also ensures customer privacy is upheld, by discovering what features are correlated with identifiable information so they can be removed.	[233]
[...] explainable-AI calls for [...] privacy [...].	[236]
Revealing the hidden information of stakeholders is in conflict with secrecy practices [...]	[241]
Concerning data transparency, it is also important to know whether it reveals any identity, self (i.e., personal), or hidden information [...]. Revealing identity information can diminish, if not demolish, the anonymity of stakeholders [...].	[241]
This is where transparency and privacy intersect and transparency may threaten stakeholders' privacy [...].	[241]
[...] it can threaten both privacy and security, even though transparency is seen as a positive concept [...].	[241]
As revealing self information can endanger stakeholders' privacy requirements, it must be ensured at early stages of system analysis that the revealed data complies with privacy regulations [...].	[241]
This should be noted that transparency is not the opposite of privacy, but there are occasions where the two concepts get at odds with each other, leading to conflicting demands between transparency and privacy [...].	[241]
Transparency could be even twinned with privacy and data protection in the sense of being transparent about the regulations about the right to hide or the obligation to reveal information [...].	[241]
However, transparency may also have negative consequences, e.g. regarding privacy [...].	[391]
Privacy is a key concern if information could become available to "adversaries", ie. malicious parties.	[417]
Every explainability approach should be accompanied by a critical evaluation of its privacy and security implications and a discussion about mitigating these factors.	[442]
This exchange of knowledge between the explainee and the explainability system [...] poses a significant [...] privacy risk.	[443]
A weakness of this training data based influence interpretation approach is the privacy issue of training data.	[510]
To operationalize [...] Transparency [...] one may need to disclose a large amount of information that could jeopardize Privacy.	[512]

Table 23: Quotes for the desideratum *Reliability*.

Quote	Src.
Specific concerns that require explanations include [...] reliability [...].	[1]
It has been suggested that the intelligibility of system behavior is an important factor in ensuring that the user understands how the CDSS operates [...]. This in turn could [...] ensure that the clinician forms a more accurate picture of the system's reliability.	[93]
[...] desiderata that can be optimized through interpretability: [...] Reliability [...] — Ensure that small changes in the input do not cause large changes in the prediction.	[99]
Interpretability is used to confirm other important desiderata of ML systems: [...] Properties such as reliability and robustness ascertain whether algorithms reach certain levels of performance in the face of parameter or input variation.	[155]
[...] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
Explanations can be used to help assess the reliability of systems [...].	[475]
Questions that experts in artificial intelligence (AI) ask opaque systems provide inside explanations, focused on [...] reliability [...].	[189]
Providing justifications seemed to benefit the perceived reliability regardless of gender.	[357]
This limitation is a serious roadblock for applications in which trust and reliabiltiy are critical. In order to solve this problem, researchers have begun developing techniques to [...] provide explanations as to why the agent chose a particular action [...].	[362]
[...] a system that generates an explanation that is not coherent with its output is not reliable.	[389]
[...] users who are recipients of system-generated recommendations [...] prefer information that indicate the systems prior performance, such as reliability data [...].	[482]

Table 24: Quotes for the desideratum *Reliance*.

Quote	Src.
In some cases, explanation increases [...] reliance [...].	[72]
In some cases, explanation increases [...] reliance [...], but in others an explanation may have the opposite effect if the level of detail it contains is deemed insufficient [...].	[72]
Whilst a more detailed explanation may promote over-reliance, we argue that providing no explanation at all is not a viable option [...].	[93]
[...] without providing explanations there is a danger that users will rely too much on themselves because they do not understand how the system works.	[93]
[...] a more detailed explanation may promote over-reliance [...].	[93]
[...] less detailed explanations made participants question the system's reliability and led to self-reliance problems.	[93]
[...] there is a need to explain how they work so that users and decision makers can develop appropriate [...] reliance.	[231]
Greater transparency allows people to question and critique a system in order to develop appropriate reliance [...].	[388]
[Explanations] also help users to know what the limitations of a system are, and when they can rely on it [...].	[388]

Table 25: Quotes for the desideratum *Responsibility*.

Quote	Src.
[...] counterfactuals [...] increase people's ascriptions of blame and fault to the action [...].	[95]
Counterfactuals [...] can amplify judgments of blame.	[95]
[...] since responsible inquiry on the web requires that users be aware of certain key details, the opacity of those details bars us from using the devices responsibly.	[130]
[...] society still will not demand on explanation unless the explanation can be acted on in some way. This could mean [...] assigning a blame and providing compensation for injuries caused by past decisions.	[156]
[...] this principle, which we synthesise as "explicability" [...] (as an answer to the question: "who is responsible for the way it works?") [...].	[175]
[...] explanations have several functions other than the transfer of knowledge, such as [...] assignment of blame.	[339]
Transparency and the ability to explain AI decision making are core requirements for important aspects such as [...] responsibility [...] of algorithms.	[349]
[...] "[the] human user bears the ultimate responsibility for action" and, therefore, she should be able to explain the decision.	[358]
Finally, a purely functional understanding of a model would also impede [...] public responsibility for the decisions of the model.	[365]
In AI, a user of an expert system can be held responsible for a decision made with use of the system, as long as the user has a reasonable way of knowing whether the decision proposed by the system is sensible (explanation-for-confidence).	[373]
Only if the right kind of information is given can informed consent on using the system and its outputs be established, and can responsibility be clearly allocated.	[373]
Transparency mechanisms can convey a sense of [...] individual users feeling like they are in some way responsible for the outputs of the algorithm.	[388]
The idea is that an explanation [...] will allow a human being to have meaningful control [...]—enabling the ascription of moral responsibility to that human being (or set of human beings).	[397]
It should be clear that explicability is considered to be an important part of achieving so-called [...] 'responsible' [...] AI.	[397]
An intelligent robot that is explainable yields several important advantages. [...] know where the blame lies when things go wrong.	[426]

Table 26: Quotes for the desideratum *Robustness*.

Quote	Src.
Interpretability facilitates the provision of robustness by highlighting potential adversarial perturbations that could change the prediction.	[54]
[Explanations] can help a domain expert intuitively assess how solid the prediction is.	[73]
[...] explanation-generating algorithms will be instrumental in developing more robust systems for use in [...] critical domains.	[77]
[...] desiderata that can be optimized through interpretability: [...] Robustness — Ensure that small changes in the input do not cause large changes in the prediction.	[99]
Interpretability is used to confirm other important desiderata of ML systems: [...] Properties such as reliability and robustness ascertain whether algorithms reach certain levels of performance in the face of parameter or input variation.	[155]
[...] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
To enable [...] robust [...] integration of such systems, the end users require them to support interpretability [...] in decision-making [...].	[187]
The benefits of our proposed explainability framework are three-fold. We can identify [...] fix or delete adversarial examples [...] in order to improve model robustness.	[322]

Table 27: Quotes for the desideratum *Safety*.

Quote	Src.
These explanations would [...] incite the user to understand [...] the agents, thereby improving the levels of [...] safety [...].	[30]
These explanations would [...] incite the user to understand [...] the agents, thereby [...] avoiding failures [...].	[30]
Explanations are particularly essential for intelligent systems in [...] safety-critical industry [...].	[29]
Interpretability enables MLs models to be tested, audited, and debugged, which is a path towards increasing their safety [...].	[99]
We argue that the need for interpretability stems from an incompleteness in the problem formalization [...]. Below are some illustrative scenarios: Safety: For complex tasks, the end-to-end system is almost never completely testable [...].	[155]
To enable safe [...] integration of such systems, the end users require them to support interpretability [...] in decision-making [...].	[187]
Explanation technologies are an immense help to companies for creating safer [...] products [...].	[201]
The results show a clear and consistent order of the three visualizations regarding the efficiency to increase trust [...].	[214]
[...] explainable-AI calls for [...] safety [...].	[236]
A number of motivations one might have in seeking explicability were cited earlier, such as safety [...].	[280]
Explanations can help increase [...] safety by identifying when the recommendation is reasonable and when it is not.	[288]
Examples include improving safety, where a user might use the explanation to determine when the machine learning system will make a mistake [...].	[288]
The question of what kinds of explanation a human can utilize implies the presence of a downstream task. [...] extrinsic tasks include goals such as safety – given the input, output, explanation, and observations of the world, does the explanation help the human user identify when the agent is going to make a mistake?	[350]
[...] designers of secure websites need to explain to the banking client why they can safely do their transactions online [...].	[373]
Currently there is much debate regarding the safety of [...] data processes in general, leading to investigations regarding the explainability of AI-supported decision making.	[391]
To date, many reasons have been suggested for making systems explainable [...]: [...] To explain the agent's choices to guarantee safety concerns are met [...].	[399]
Additional goals [of explanation] include [...] guaranteeing safety concerns [...]	[399]
This exchange of knowledge between the explainee and the explainability system [...] poses a significant safety [...] risk.	[443]
We list several types and goals of transparency. [...] To facilitate monitoring and testing for safety standards.	[490]

Table 28: Quotes for the desideratum *Satisfaction*.

Quote	Src.
[...] explanations can serve a multiplicity of aims, such as [...] increasing the ease of use of a system (satisfaction) [...].	[51]
Arguably, the most important contribution of explanations is [...] to allow [users] to make more informed and accurate decisions about which recommendations to utilize (satisfaction).	[70]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: [...] satisfaction [...].	[72]
Explanation has been shown to be important for user [...] satisfaction in a number of studies.	[73]
Other studies [...] have also shown that users are overwhelmingly more satisfied with systems that contain some form of justification [...].	[73]
Our approach helps users [... increase] their satisfaction with the explanation.	[74]
Other studies have also shown that users are overwhelmingly more satisfied with systems that contain some form of justification [...].	[74]
Algorithmic explanations have been found to [...] increase user satisfaction [...].	[96]
Prior work shows that explanations [...] can increase user satisfaction [...].	[96]
[...] explanatory aim (e.g., [...] user satisfaction [...]).	[110]
content-based tag cloud explanations are more effective and helpful to increase users' satisfaction [...].	[110]
[...] explanation purpose (e.g., [...] user satisfaction [...])[...].	[111]
User studies showed that this [explanation method] is helpful to [...] increase [user] satisfaction with the system	[111]
The perceptions arising from the use of explanation facilities include [...] user satisfaction [...].	[137]
[...] user satisfaction [... are] likely benefits of explanation facilities for recommender systems.	[137]
[...] the interfaces with explanations have a positive effect on understandability, which then has a positive effect on satisfaction [...].	[151]
"[...] users are ""more satisfied with explanation facilities which provide justifications for the recommendations""."	[151]
[...] explainable interfaces also have a positive effect on time, that also has a positive effect on satisfaction.	[151]
[...] explanations of recommendations in the image domain are useful and increase user satisfaction [...].	[152]
AI rationalization has a number of potential benefits over other explainability techniques: [...] humanlike communication [...] may afford [...] advantages such as higher degrees of satisfaction [...]	[166]
[...] transparency of music recommendations increased participants' satisfaction with the recommendation [...].	[173]
[...] good explanations can [...] increase user satisfaction [...].	[198]
Explanations that conform to Toulmin's model should be more persuasive [...]. Thus, they should lead to greater [...] satisfaction [...].	[199]
[...] satisfaction [... was] strongly affected by the presence of justification-type explanations [...].	[199]
[...] explanations will help properly set expectations [...]. Setting customer expectations correctly [...] is important to their satisfaction with the system.	[225]
The justifications have a significant impact on perception of [...] self-anticipated satisfaction with the system.	[275]
The proposed concept performed significantly better compared to [...] simple explanations in terms of our [...] side goals to increasing perceived [...] satisfaction.	[291]
[...] we discuss how [...] providing explanations [...] has the potential to increase user satisfaction [...].	[305]
The importance of [...] explanation on improving user satisfaction [...] has been extensively discussed.	[348]
[...] motivations for explanations in recommender systems: [...] satisfaction, which increases users' willingness to continue use.	[348]
Looking at the historical developments, we can observe [...] that other potential purposes [of explanations], like user satisfaction [...] received more attention in the recent past.	[358]
Explanation Purposes Identified in Primary Studies [...]: Satisfaction. [...] Increase the ease of use or enjoyment.	[358]
Explanations that help users understand how a system works have demonstrated a positive relationship with user satisfaction with the system [...].	[388]
Sevastjanova, Rita, et al. - Going beyond Visualization - Verbalization as Complementary Medium to Explain Machine Learning Models (2018)	[422]

Continued on next page ...

Table 28 – continued from previous page

Quote	Src.
"We examine the results in an attempt to ascertain whether explanation can be correlated with improved performance and/or user satisfaction; Our earlier results demonstrated clearly that the use of argumentation-based dialogue improves both of these measures"	[436]
We show several examples of explanations and ask participants to judge the examples on four [...] dimensions: [...] satisfaction [...].	[229]
[...] possible aims when explaining the outcomes of an algorithm to users: [...] satisfaction.	[229]
Among other things, good explanations could [...] increase satisfaction [...].	[463]
Among other things, good explanations could [...] increase satisfaction [...].	[464]
Possible aims for explanations: Satisfaction. Increase the ease of usability or enjoyment	[464]
Explanations may help users enjoy movies more, rather than serve merely as decision aids.	[464]
Explanations can also serve other aims such as [...] increase satisfaction [...].	[465]
[...] in our work we have found that [...] personalized explanations may lead to greater user satisfaction [...].	[465]
[...] one can measure how understandable an explanation is, which can contribute to e.g. [...] satisfaction [...].	[465]
Explanatory criteria and their definitions: Satisfaction [...]. Increase the ease of use or enjoyment.	[465]
Explanations can increase satisfaction by clarifying or hinting that the system considers changes in the user's preferences.	[465]
Cramer et al. have investigated the effects of transparency on other evaluation criteria such as [...] satisfaction [...].	[465]
Explanations have been found to increase user satisfaction with [...] the overall recommender system [...].	[465]
Explanatory aims: Satisfaction. Increase the ease of use or enjoyment.	[466]
[...] explanations can be more focused on helping users make decisions (about the items) that they are happy with [...]	[466]
[...] providing an explaining icon [...] plays a crucial role in contributing to the factor of user satisfaction.	[473]
Explanations provide many benefits, from improving user satisfaction to helping users make better decisions.	[478]
[...] recent work suggests a broader set of goals including trust, user satisfaction, and transparency [...].	[478]
[...] the most important contribution of explanations is [...] to allow [users] to make more informed and accurate decisions about which results to utilize (i.e., satisfaction) [...].	[487]

Table 29: Quotes for the desideratum *Science*.

Quote	Src.
Asking for explanations is a helpful tool to learn new facts, to gather information and thus to gain knowledge. [...] It will come as no surprise if, in future, XAI models taught us about new and hidden laws in biology, chemistry and physics.	[4]
[...] desiderata that can be optimized through interpretability: [...] Causality — Ensure that only causal relationships are picked up.	[99]
Possible incentives for asking a plain fact question, are sheer intellectual curiosity, and a desire to causally connect object a having property P to events with which we are more familiar.	[497]
Interpretability is used to confirm other important desiderata of ML systems: [...] Causality implies that the predicted change in output due to a perturbation will occur in the real system.	[155]
[...] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
We argue that the need for interpretability stems from an incompleteness in the problem formalization [...]. Below are some illustrative scenarios: Scientic Understanding: The human's goal is to gain knowledge. [...] the best we can do is ask for explanations we can convert into knowledge.	[155]
[...] in some application domains users need to understand the system's recommendations enough to legally explain the reason for their decisions to other people.	[179]
The main purpose to provide explainability of a model also varies, e.g., the goal might be to support [...] causality [...].	[234]
The use of ML algorithms by scientists for the purpose of generating causal explanations is one such case.	[280]
Algorithms whose inner workings are not interpretable will not enable us to produce causal explanations of the world.	[280]
Ratti and López-Rubio [...] have emphasized a connection between interpretability and causal explanation in molecular biology.	[280]

Continued on next page ...

Table 29 – continued from previous page

Quote	Src.
[...] explanations could help [...] generate hypothesis about the causality.	[304]
[...] it could be that explanations [...] play a role in discovery and confirmation, which in turn produces theories that support future prediction and intervention.	[308]
[...] explanations in these cases can be helpful for knowledge discovery [...].	[394]
[...] an agent that provides an explanation for its decision might further human understanding of a medical phenomenon.	[394]
Explicable AI may be extremely valuable to researchers and others who would be able to use explanations to better understand their domain.	[397]
Explanations geared beyond the immediate user can also be those geared for researchers to help facilitate scientific knowledge discovery [...].	[399]
Additional goals [of explanation] include [...] knowledge/scientific discovery.	[399]
To date, many reasons have been suggested for making systems explainable [...]: [...] Knowledge/scientific discovery [...].	[399]
[...] reasons why ML interpretability is desired, namely [...] causality [...].	[417]
Explainable artificial intelligence presented [...] mentions [...] discover.	[417]
Interpretability of the machine learning predictions is important for a variety of reasons. Scientists [...] may be driven by a scientific curiosity hoping to use machine learning to elicit new knowledge from data.	[441]
Most [explainability methods] are designed for transparency: explaining [...] ([...] to [...] elicit knowledge form a predictive model or the data used to build it, or extract a causal relation).	[442]
We can expect more visualizations of this type to be created and used with other kinds of computational models in systems biology.	[454]
[...] visualization can itself be a knowledge generator as it intuitively leads the analyst from observed model outcomes to potential hypothesis about the observed data.	[476]
Interpretation in this case is used to acquire new knowledge through visualization.	[476]
[...] we further argue that XAI should support abductive and [hypothetico-deductive] reasoning i.e., in addition to providing counterfactuals to help users find causes, we should provide explanations to allow users to generate and test hypotheses to further narrow down potential causes.	[488]
Explanations describe the decision made by a machine learning model [...] for an increase in insight to the domain area for instance uncovering causality [...].	[485]
Theoretically, the Black Box Problem makes it difficult to evaluate the potential similarity between artificial neural networks and biological brains [...]. [...] Investigators within the Explainable Artificial Intelligence [...] research program intend to ward off these consequences [...].	[503]

Table 30: Quotes for the desideratum *Security*.

Quote	Src.
[...] cyptography brings some transparency problems, because they can improve security [...] by bringing lack of transparency [...]	[128]
[...] explainable-AI calls for [...] security [...].	[236]
Similarly, security and transparency are sometimes viewed as two antagonistic requirements [...].	[241]
[... transpanency] can threaten both privacy and security, even though transparency is seen as a positive concept [...].	[241]
However, explainability requirements may conflict with other softgoals such as [...] security.	[277]
Whether transparency also contributes to the security of the system itself is heavily debated: some would argue that making the protection mechanisms public will enhance the capabilities of the attackers [...].	[373]
Explanations are thought to bridge the gap between ‘actual security’ and ‘perceived security’.	[373]
Transparency is [...] considered essential for allowing the users to understand what the designers have done to protect them.	[373]
Whether transparency also contributes to the security of the system itself is heavily debated: some [...] would argue that protection mechanisms can be improved by public scrutiny.	[373]
"[...] the explanation must be provided for a given user while also considering the implications on the system’s security goals."	[399]

Continued on next page ...

Table 30 – continued from previous page

Quote	Src.
Every explainability approach should be accompanied by a critical evaluation of its privacy and security implications and a discussion about mitigating these factors.	[442]
Another example can be a security trade-off between ante-hoc [explainability] approaches that reveal information about the predictive model itself [. . .].	[442]
This exchange of knowledge between the explainee and the explainability system [. . .] poses a significant [. . .] security [. . .] risk.	[443]
[. . .] non-expert users will need an explanation [. . .] that also teaches them how to use the system correctly and securely [. . .].	[479]
Legally, opacity prevents regulatory bodies from determining whether a particular system processes data fairly and securely [. . .]. [. . .] Investigators within the Explainable Artificial Intelligence [. . .] research program intend to ward off these consequences [. . .].	[503]

Table 31: Quotes for the desideratum *Transferability*.

Quote	Src.
[. . .] purposes motivating the need for interpretable AI models, such as [. . .] transferability [. . .].	[54]
Explainability is also an advocate for transferability, since it may ease the task of elucidating the boundaries that might affect a model, allowing for a better understanding and implementation.	[54]
The explanations lead to insights of feature transfer for users without ML expertise, and in turn allow them to further improve a transfer learning approach with more optimized settings.	[109]
Answering plain fact questions can also serve other purposes [. . .]. One such purpose is related to the development of a new program by using components of an outdated program.	[497]
If our engineer prefers creating a new program [. . .], and he wants to reuse [. . .] components [. . .], he has to know which [. . . fulifll] this function. The corresponding knowledge-seeking question can be reformulated into the plain fact explanation-seeking question.	[497]
The main purpose to provide explainability of a model also varies, e.g., the goal might be to support [. . .] transferability [. . .].	[234]
[. . .] reasons why ML interpretability is desired, namely [. . .] transferability [. . .].	[417]

Table 32: Quotes for the desideratum *Transparency*.

Quote	Src.
There has been increased attention into [. . .] transparent algorithms [. . .], with [. . .] DARPA's Explainable AI (XAI) initiative [. . .].	[1]
[. . .] eXplainable Artificial Intelligence (XAI) emerged with the aims of fostering transparency [. . .].	[30]
[. . .] transparency [. . . is] among the listed motivations for the explanations.	[30]
Explanations are particularly essential [. . .] as it [sic!] raises [. . .] transparency in the system.	[29]
Explainable recommendation methods [. . .] improve [. . .] transparency [. . .].	[43]
Explanations can serve a multiplicity of aims, including transparency [. . .].	[51]
Indeed, one of the aims that explanations can serve is to provide transparency [. . .].	[51]
A system's ability to explain its recommendations [. . .] makes its reasoning more transparent [. . .].	[70]
[. . .] explanation approaches might serve regulatory goals of rendering algorithmic decision-making more [. . .] transparent.	[72]
The potential for [. . .] explanation systems to provide justice-related information, fulfilling the policy goals of transparency [. . .], has recently been noted [. . .].	[72]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: transparency [. . .]	[72]
While considering the integration of explanations in a system, the goal may be to add transparency [. . .] but it may result in the opposite effects.	[108]
While considering the integration of explanations in a system, the goal may be to add transparency [. . .].	[108]

Continued on next page . . .

Table 32 – continued from previous page

Quote	Src.
Explanations may also benefit the auditability of a system, mainly related to its technical aspects and data transparency.	[108]
[...] explanation can be [...] for increasing system transparency [...].	[110]
The importance of explanation interfaces in providing system transparency [...] has been recognized in a number of fields [...]	[110]
[...] explanatory aim (e.g., increasing system transparency [...]).	[110]
[...] explanations can significantly increase [...] perceived recommendation transparency [...].	[111]
[Users] perceived the system to be significantly better at [...] transparent and good recommendations.	[111]
[...] explanation can be [...] for increasing system transparency [...].	[111]
The importance of explanation interfaces in providing system transparency [...] in a number of fields [...].	[111]
[...] explanation purpose (e.g., increasing system transparency [...]) [...]	[111]
Such a classification model offers transparency: each prediction can be explained trivially by analyzing the terms that were present in the text [...].	[118]
[...] explanations can increase transparency [...].	[386]
[...] explanations can be useful to expose the inner workings of an AI system to its users, thus fostering transparency [...].	[386]
[...] transparency [... are] likely benefits of explanation facilities for recommender systems.	[137]
Explainable AI (XAI) is a field broadly concerned with making AI systems more transparent [...]	[150]
This lack of transparency drives a sweeping call for explainable artiicial intelligence (XAI) in industry, academia, and public regulation.	[150]
Numerous studies have demonstrated the benefits of explanations in intelligent systems [...], including [...] increasing transparency of system reasoning [...].	[165]
[...] some approaches for increasing the transparency of filtering process [...] involve explanations [...].	[165]
The need for explainable AI is motivated mainly by three reasons: the need for trust [...].	[177]
[...] an explainable system may be transparent [...].	[189]
Explanations, by virtue of making the performance of a system transparent to its users [...].	[199]
[...] the recommendation process is transparent by explaining its recommendations [...].	[215]
An explanation behind the reasoning of a ACF recommendation provides transparency into the workings of the ACF system.	[222]
Building an explanation facility into a recommender system [...] removes the black box from around the recommender system, and provides transparency.	[222]
We present work in progress on explainability to support transparency in human AI interaction.	[234]
[...] explainability provides transparency [...].	[234]
The presence of explanations assisted participants in the conception of an accurate mental model, increasing perceived transparency [...]	[235]
[...] the same explanation [...] with every HR request [...] ([...] may cause unnecessary transparency).	[241]
The proposed concept performed significantly better compared to [...] simple explanations in terms of our main goals to increase transparency [...].	[291]
[...] outcome explanation, the input and output visualization, which made everyone's preferences and outcomes transparent [...].	[302]
Explainability is the first step in achieving a transparency goal [...].	[322]
[...] we introduced Bart, an algorithm for jointly personalizing recommendations and associated explanations for providing more transparent [...] suggestions to users.	[326]
[...] motivations for explanations in recommender systems: [...] transparency [...].	[348]
[...] the most common explanation purpose is to provide transparency [...].	[358]
[...] the suggestions [...] must be perceived to be [...] transparent [...], and explanations are key to this.	[358]
Explanation Purposes Identified in Primary Studies [...]: Transparency. [...] Explain how the system works.	[358]
The main goal of Explainable Artificial Intelligence (XAI) has been variously described as a search for explainability, transparency and interpretability [...].	[365]
[...] explanations of what procedures are built into the design and what procedures exist if something goes wrong would then contribute to transparency.	[373]

Continued on next page ...

Table 32 – continued from previous page

Quote	Src.
The importance of explanation interfaces in providing system transparency [...] has been well recognized in a number of fields	[382]
The importance of explanation interfaces in providing system transparency [...] has been well recognized in a number of fields [...].	[383]
Furthermore, if the agent makes a mistake, the generated reasons could provide transparency [...].	[394]
An intelligent robot that is explainable yields several important advantages. [...] making the decision processes transparent [...].	[426]
To be trusted, a system has to demonstrate [...] that the process leading to the decision is transparent and accountable [...]. Explanations form a vital part of satisfying these requirements.	[427]
Every explainability approach should be accompanied by a list of its intended applications [...]. Most of them are designed for transparency [...].	[442]
[...] an explanation in the form of a reasoning trace [...] would be presented. This would offer the user a degree of transparency into how the system reached its conclusions.	[444]
We show several examples of explanations and ask participants to judge the examples on four [...] dimensions: transpareny [...].	[229]
[...] possible aims when explaining the outcomes of an algorithm to users: transparency [...].	[229]
Possible aims for explanations: Transparency. Explain how the system works.	[464]
Explanations can provide that transparency, exposing the reasoning and data behind a recommendation.	[465]
In this way, we distinguish between different explanation such as e.g. explaining the way the recommendation engine works (transparency) [...].	[465]
Explanatory criteria and their definitions: Transparency [...]. Explain how the system works.	[465]
[...] explanations can provide transparency, exposing the reasoning and data behind a recommendation.	[466]
Explanatory aims: Transparency. Explain how the system works.	[466]
[...] recent work suggests a broader set of goals including trust, user satisfaction, and transparency [...].	[478]
Pieters also discusses two main goals that an explanation may have: transparency (e.g., to allow users to understand what the designers have done to protect them) [...].	[479]
Explanations are provided to support transparency, where users can see some aspects of the inner state or functionality of the AI system.	[485]
Abdul had identified other goals for XAI, such as providing transparency [...].	[485]
Investigators within the Explainable Artificial Intelligence [...] research program [...] use [...] analytic techniques capable of rendering opaque computing systems transparent.	[503]
It is generally agreed upon that the goal of XAI is to [...] improve the transparency of the system [...].	[511]

Table 33: Quotes for the desideratum *Trust*.

Quote	Src.
Specific concerns that require explanations include [...] trust [...].	[1]
[...] people should be able to [...] trust [the technology...]. [...] prior work has identified issues [...] when this is not the case. [...] To address these problems, machine learning algorithms need to be able to explain how they arrive at their decisions.	[1]
XAI aims to "enable human users to [...] appropriately trust [...] the emerging generation of artificially intelligent partners".	[4]
Using XAI systems [...] ensures that there is [... a] way to defend algorithmic decisions as being fair and ethical, which leads to building trust.	[4]
These explanations would [...] incite the user to understand [...] the agents, thereby improving the levels of trust [...].	[30]
Increasing user's trust in the system [... is] among the listed motivations for the explanations.	[30]
Explanations are particularly essential [...] as it [sic!] raises trust [...] in the system.	[29]
Intelligent systems that are explaining their decisions to increase the user's trust [...] are widely studied.	[29]
[...] explanation and interpretation are required to enable domain experts and users to [...] trust [...] the [...] model [...].	[43]

Continued on next page ...

Table 33 – continued from previous page

Quote	Src.
Explainable recommendation methods [...] improve [...] user trust [...].	[43]
[...] explainability will also enhance trust in the system at the level of the users [...]	[43]
[Explainable Artificial Intelligence] aims at producing intelligent systems that reinforce the trust of the users [...].	[48]
[...] explanations can serve a multiplicity of aims, such as inspiring the user's confidence in the system (trust) [...].	[51]
The approach we present to increase trust in these systems consists in providing the user with semantic and context-dependent information as well as logical information about plans.	[69]
In some cases, explanation increases trust [...].	[72]
Tintarev and Masthoff identify seven purposes for recommender system explanations, namely: [...] trust [...].	[72]
In some cases, explanation increases trust [...], but in others an explanation may have the opposite effect if the level of detail it contains is deemed insufficient [...].	[72]
Later studies [...] showed that explanations significantly increase users' [...] trust [...].	[73]
In order to engender trust in AI, [...] the underlying AI process must produce justifications and explanations that are both transparent and comprehensible to the user.	[77]
Lee and See provide an extensive description of the need for and methods to achieve user trust in a computer system: comprehensible information.	[82]
[...] giving a fuller explanation of the facts used in making a diagnosis had a positive effect on trust [...]	[93]
[...] XAI may have the overall goal of improving trust in decisions made by AI systems [...].	[95]
Careful constraints on counterfactuals are required to provide interpretable models of the decisions of AI systems that people can trust [...].	[95]
To increase trust in such systems by human users, [...] there is a need to enable AI systems to provide [...] explanations [...].	[95]
[...] comparative explanations could help establish a more appropriate level of trust.	[96]
Algorithmic explanations have been found to [...] increase user [...] trust [...].	[96]
Prior work shows that explanations [...] can increase user [...] trust [...].	[96]
However, some found that too much explanation can create confusion and degrade trust [...].	[96]
[...] explaining a ML model's decisions provides a way to check [...] trust.	[99]
desiderata that can be optimized through interpretability: [...] Trust	[99]
[...] medical students [...] trusted the system more when presented with an explanation [...].	[101]
We can see that most people seem to agree that the explanations were helpful and easy to understand. In fact, the majority of people strongly agreed that their trust of the robot increased during the study [...].	[103]
[...] explanatory aim (e.g., [...] user trust [...]).	[110]
[...] explanation purpose (e.g., [...] user trust [...])[...].	[111]
[...] explanation can cause users to overestimate item quality, which may lead to mistrust [...].	[111]
[...] trust of a personalized mobile application depends highly on perceived information transparency [...].	[112]
[...] designers may [...] increase system transparency to breed trust [...].	[112]
[...] XAI will be key for both expert and non-expert users to enable them to have [...] appropriate level of trust [...].	[116]
Understanding the reason behind a classification allows us to establish trust in further predictions [...].	[118]
[...] explanations can be useful to expose the inner workings of an AI system to its users, thus fostering [...] trust [...].	[386]
"[...] "how" explanations open the black box and instill trust by enabling users to verify that an algorithm has accurately and fairly ""produc[ed] and certif[ied] knowledge""."	[123]
[...] explanations allow users to make inferences about a system's abilities and underlying motives, which form the basis of [...] trust in a system.	[123]
Systems researchers have emphasized the importance of explanations as a means of influencing user [...] trust in systems by increasing confidence in systems' abilities [...].	[123]
Increasing transparency of a system can help users decide whether they can trust the system.	[124]
Increasing transparency of user-adaptive systems could thus increase trust [...] of such systems [...].	[124]
[...] transparency in many cases may be essential to assure trust.	[128]
[...] self-driving cars, which can demonstrate transparency in operations, will help promote trust [...].	[127]
[Having transparency] would [...] increase trust [...].	[127]

Continued on next page ...

Table 33 – continued from previous page

Quote	Src.
The lack of explainability results in the lack of trust [. . .].	[131]
Explainability is [. . .] a pre-requisite for practitioner trust [. . .].	[131]
If software practitioners do not understand a model's predictions, they would not blindly trust those predictions, nor commit project resources or time to act on those predictions.	[131]
We have argued for explainable software analytics to facilitate human understanding of machine prediction as a key to warrant trust from software practitioners.	[131]
The perceptions arising from the use of explanation facilities include trust [. . .].	[137]
People gather more reliable information about abilities and mental states. As a result, people will correctly calibrate their trust in such systems.	[196]
When people trust the explanation, it follows that they would be more likely to trust the underlying ML systems.	[150]
Explainable AI (XAI) is a field broadly concerned with making AI systems more transparent so people can confidently trust an AI system [. . .].	[150]
Both effects are mediated by understandability, which could mean that users only trust something they understand	[151]
Interpretability is used to confirm other important desiderata of ML systems: [. . .] trusted systems have the confidence of human users.	[155]
[. . .] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
By exposing the logic behind a decision, explanation can be used to prevent errors and increase trust.	[156]
[. . .] access to an explanation might decrease observers' trust in some decisions [. . .].	[156]
We find that including certain transparency features [. . .] does improve user trust [. . .].	[161]
Numerous studies have demonstrated the benefits of explanations in intelligent systems [. . .], including building user trust [. . .].	[165]
Studies have explored [. . .] benefits related to providing explanations [. . .] many focusing on the advantage of earning users' trust [. . .] in the systems [. . .].	[165]
[. . .] some approaches for increasing [. . .] the users' trust in recommender systems involve explanations [. . .].	[165]
[. . .] participants preferred rationales with transparency so that they can [. . .] trust the robot in a situation where expectations are violated.	[167]
[. . .] some studies [. . .] reveal how transparency can also undermine trust [. . .].	[173]
[. . .] studies [. . .] tend to find positive outcomes for organizations, for example, positive effects on organizational trust [. . .].	[173]
[. . .] explicability is a critical tool to build public trust in [. . .] the technology.	[175]
The need for explainable AI is motivated mainly by three reasons: the need for trust [. . .].	[177]
First, understanding a computer-induced model is often a prerequisite for users to trust the model's predictions [. . .].	[179]
[. . .] explanations [. . .] promote objectives such as trust [. . .].	[180]
In order for humans to trust black-box methods, we need explainability [. . .].	[188]
These outside explanations can be used to build trust [. . .].	[189]
[. . .] transparency is particularly useful in building trust in a system [. . .].	[190]
[. . .] transparency can provide a building block on which to establish trust in systems for which no other basis already exists [. . .].	[190]
[. . .] the use of complex explanation systems can address the majority of the trust concerns [. . .].	[190]
Several users reported that explanations [. . .] would enable them to trust results without the need for extensive further verification.	[190]
[. . .] good explanations can help inspire trust in a recommender [. . .].	[198]
Explanations, by virtue of making the performance of a system transparent to its users, are influential [. . .] for improving users' trust [. . .].	[199]
Possibly one would feel more comfortable and trusting of an agent if it is able to explain what it is doing and why.	[199]
Explanations that conform to Toulmin's model should be more persuasive [. . .]. Thus, they should lead to greater trust [. . .].	[199]
Explanation technologies are an immense help to companies for creating [. . .] more trustable products [. . .].	[201]
Another property that influences the trust level of a model is usability: people tend to trust more models providing information that rassist them to acocmplish a task [. . .].	[201]

Continued on next page . . .

Table 33 – continued from previous page

Quote	Src.
The results show a clear and consistent order of the three visualizations regarding the efficiency to increase [...] perceived safety [...].	[214]
To improve [...] trust, transparency and explainability become increasingly important for practical recommender systems.	[215]
[...] explanations [...] can play a key role [...] to enhance trust in the system [...].	[221]
The explanations can be useful to enhance his trust in the classifier [...].	[221]
Explanation capabilities provide a solution to building trust [...].	[222]
Users will be more likely to trust a recommendation when they know the reasons behind that recommendation.	[222]
[...] there is a need to explain how they work so that users and decision makers can develop appropriate trust [...].	[231]
Using human trust as a metric of evaluation for the effectiveness of explanations has also been studied [...].	[233]
[...] explainability [...] contributes to trust [...].	[234]
The main purpose to provide explainability of a model also varies, e.g., the goal might be to support trust [...].	[234]
The trust levels of participants in the with-explanation group were influenced strongly by the perceived ability of the AutoCoder. [...] Without the presence of explanations, [...] their level of trust did not increase.	[235]
[...] explainability of AI could help to enhance trust of medical professionals in future AI systems.	[236]
[...] explainable-AI calls for [...] trust [...].	[236]
Such a shift in transparency provision can lead to positive side effects, such as more trust.	[241]
[...] transparent design of algorithmic interfaces can promote awareness and foster trust [...].	[274]
Transparency may promote [...] users' trust in a system by changing beliefs about its trustworthiness.	[274]
However, providing too much information eroded this trust [...].	[274]
Transparency may [...] erode users' trust in a system by changing beliefs about its trustworthiness	[274]
The justifications have a significant impact on perception of [...] trust [...].	[275]
Explanations enable understanding and thereby foster trust [...].	[277]
If algorithms lack transparency, domain experts or the public will not trust them [...].	[280]
Likewise, there is the case of public trust. To the extent that public acceptance of ML algorithms requires that end users have some grasp of the inner workings [...], the notion of interpretation acquires heightened importance.	[280]
Such summaries have been shown to [...] facilitate trust [...].	[289]
Explanations can help increase trust [...] by identifying when the recommendation is reasonable and when it is not.	[288]
Examples include [...] increasing trust, where a user might be convinced to use a machine learning system if it justifies its actions in plausible ways.	[288]
"Explanation is often embraced as a cure for ""black box"" models to gain trust and adoption."	[304]
[...] generating explanations of application behavior [...] has been employed [...] with the goal of increasing user trust [...].	[305]
We aim to evaluate the effect of explaining model outputs, specifically large errors, on users' attitudes towards trusting [...].	[312]
An appropriate explanation can promote trust in the system [...].	[316]
Driven by lack of trust [...] , there are many calls for Artificial Intelligence (AI) systems to become more transparent, interpretable and explainable.	[316]
Explainable Machine Learning (XAI) [...] enables human users to [...] appropriately trust [...] emerging generation of artificially intelligent partners.	[322]
[...] XAI is a key part of applying ethics to AI because it increases trust in model decisions [...].	[322]
[...] a transparency goal [...] can be achieved by checking systems explanations to determine whether they satisfy desirable trust criteria.	[322]
[...] building machines that are able to explain and be explained to facilitates humans to gradually build trust [...].	[337]
The running hypothesis is that by building more [...] explainable systems, users will be better equipped to understand and therefore trust the intelligent agents [...].	[339]
[...] providing simpler explanations [...] may increase trust better than giving more likely explanations.	[339]
Transparency disclosures by data processors and controllers may prove crucial in the future to maintain a trusting relationship with data subjects.	[342]
Explanations can be necessary to [...] enhance the trust between individuals subject to a decision and the system itself [...].	[341]
The proposed visualization helps to [...] increase the users' trust in the system.	[348]

Continued on next page ...

Table 33 – continued from previous page

Quote	Src.
The importance of [...] explanation on improving user satisfaction (e.g., acceptance, trust) has been extensively discussed.	[348]
[...] motivations for explanations in recommender systems: [...] trust [...].	[348]
[...] explainability will also enhance trust at the user side [...].	[349]
Transparency and the ability to explain AI decision making are core requirements for important aspects such as trust [...].	[349]
The question of what kinds of explanation a human can utilize implies the presence of a downstream task. [...] extrinsic tasks include goals such as [...] trust [...].	[350]
[...] we present our work on using explanations to maintain the trust relationship between human and computer [...].	[357]
[...] explanation facilities have been widely investigated as a means of establishing trust in these systems since the early years of expert systems.	[358]
Transparency is also seen as key for users to develop trust toward the system [...].	[358]
[...] automatically generated explanations have been considered as a fundamental mechanism to increase user trust in suggestions made by the system.	[358]
[...] trust-building is explicitly mentioned as the goal of the explanations.	[358]
Explanation Purposes Identified in Primary Studies [...]: Trust. [...] Increase users' confidence in the system.	[358]
This limitation is a serious roadblock for applications in which trust and reliabiltiy are critical. In order to solve this problem, researchers have begun developing techniques to [...] provide explanations as to why the agent chose a particular action [...].	[362]
The main goal of Explainable Artificial Intelligence (XAI) has been variously described as [...] generating trust in the model and its predictive performance.	[365]
Several authors have argued that post hoc interpretability [...] is a necessary condition for trust [...].	[365]
[...] minimum explanations can potentially harm, but not improve user trust.	[366]
Artificial agents need to explain their decision to the user in order to gain trust [...].	[373]
Depending on the goal, an explanation can [...] aim at acquiring [...] trust.	[373]
[...] explanations may be an important prerequisite for the building of e-trust.	[373]
[...] bad explanation-for-trust may fail to create trust, and even lead to distrust.	[373]
[...] we have shown that explanation interfaces have the greatest potential to build a competence-inspired trust relationship with its users.	[382]
Participants on average built more trust in the organization-based explanation interface [...].	[383]
Transparency is also important in order to build human trust in systems.	[389]
Currently there is much debate regarding the [...] trust in data processes in general, leading to investigations regarding the explainability of AI-supported decision making.	[391]
The aim of local explanations is to strengthen the confidence and trust of users.	[391]
[...] explaining individual predictions is important in assessing trust.	[393]
Our experiments demonstrated that explanations are useful for a variety of models in trust-related tasks in the text and image domains [...]: [...] assessing trust [...].	[393]
[...] in many, if not most, cases, the explanation is beneficial [...] to forster better trust [...].	[394]
Interpretability is sometines beneficial to instill feelings of trust and understanding within the system's users.	[394]
[...] human-like rationales, despite being true reflections of the internal processes of a black-box intelligent system, promote feelings of trust [...].	[395]
The ability to help people understand their decisions through explanations or other means accessible to nonexperts will provide people with greater sense of trust [...].	[395]
Explainability can be important for other reasons, including building trust between the user and system [...].	[399]
In both cases, the information the agent provides should build trust [...].	[399]
To date, many reasons have been suggested for making systems explainable [...]: [...] To build trust in the agent's choices [...].	[399]
Additional goals [of explanation] include [...] guaranteeing [...] trust [...].	[399]
Explanations have various effects on users [...]. They can help gain users' trust [...].	[410]
[...] explanations show a significant improvement in user trust [...].	[412]

Continued on next page ...

Table 33 – continued from previous page

Quote	Src.
Participants that considered themselves very familiar with the task domain reported higher than average trust in the Dining Guru [...]. Presenting explanations [...] in some cases led these users to automation bias.	[415]
[...] reasons why ML interpretability is desired, namely trust [...]	[417]
[...] insights into the inner workings of a trained model allow users and analysts [...] to [...] gain trust in the systems they inform.	[422]
An intelligent robot that is explainable yields several important advantages. Trust. Humans tend to trust systems that they understand – or at least believe that they understand.	[426]
To be trusted, a system has to demonstrate competence [...], honesty [...] and alignment [...]. Explanations form a vital part of satisfying these requirements.	[427]
[... explanation] techniques [...] attempt to analyse the behaviour of the network in a black-box fashion for the purpose of increasing trust in the system.	[427]
Interpretability of the machine learning predictions is important for a variety of reasons. [...] in some applications [...] the users have to trust the predictions [...].	[441]
[...] some [explainability methods] can also be used to assess accountability of the underlying predictive model, e.g., debug and diagnose it to engender trust [...].	[442]
By large, explainability can improve users' trust in a predictive system [...].	[442]
[...] computing separate explanations [...] for each agent can result in situations where the explanations [...] are not consistent [...]. In the case of multiple teammates being explained to, this may cause confusion and loss of trust.	[451]
Such techniques [of explanation] can thus be essential contributors to the dynamics of trust and teamwork in human-agent collaborations by significantly lowering the communication overhead between agents [...].	[451]
Trust in a system is developed not only by the quality of its results, but also by clear description of how they were derived.	[456]
We show several examples of explanations and ask participants to judge the examples on four [...] dimensions: [...] trust [...].	[229]
[...] possible aims when explaining the outcomes of an algorithm to users: [...] trust [...].	[229]
Explanations can have many advantages, [...] inspiring user trust [...].	[463]
Among other things, good explanations could help inspire user trust and loyalty [...].	[463]
Among other things, good explanations could help inspire user trust and loyalty [...].	[464]
Possible aims for explanations: Trust. Increase users' confidence in the system.	[464]
Explanations can also serve other aims such as helping to inspire user trust and loyalty [...].	[465]
[...] one can measure how understandable an explanation is, which can contribute to e.g. user trust [...].	[465]
Explanatory criteria and their definitions: Trust. Increase users' confidence in the system.	[465]
Cramer et al. have investigated the effects of transparency on other evaluation criteria such as trust [...].	[465]
Trust is sometimes linked with transparency: previous studies indicate that transparency [...] increases user trust [...].	[465]
Explanatory aims: Trust. Increase users' confidence in the system.	[466]
[...] recent work suggests a broader set of goals including trust, user satisfaction, and transparency [...].	[478]
Research shows that explanations [...] increase trust in the recommender system [...].	[478]
He contrasts explanation-for-trust (i.e., explanation of how a system works, by revealing details of its internal operations) with explanation-for-conficence [...].	[479]
[...] non-expert users will need an explanation that increases their confidence and trust [...].	[479]
Finally, explanations are often proposed to improve trust in the system [...].	[485]
[...] explanations can improve trust because of increased user understanding [...].	[485]
We selected [...] biases for which we identify how XAI can play a role to mitigate them, and hence improve [...] trust.	[485]
To enable end users to [...] trust [...] their intelligent partners, [...] researchers have produced many [...] algorithm visualizations, interfaces and toolkits [...].	[485]
[...] they initially found that Why and Why Not explanations were most effective in promoting [...] trust.	[485]
[...] we developed five explanation strategies to [...] moderate trust.	[485]
Finally, explanations are often proposed to [...] moderate trust to an appropriate level [...].	[485]
[...] showing explanations of low confidence can further decrease this trust [...].	[485]

Continued on next page ...

Table 33 – continued from previous page

Quote	Src.
[...] in order [...] to adopt the system's customized results, he/she needs to first build trust over the system. Explaining the automatically generated recommendations would bridge the gap.	[487]
Explanations describe the decision made by a machine learning model in order to gain user [...] trust [...].	[488]
We contend explaining why a recommended article is relevant will [...] possibly foster user trust.	[489]
We list several types and goals of transparency. [...] For a user, to [...] build a sense of trust in the technology.	[490]
Human user needs to know why (or why not) an AI agent makes such decisions, so he can trust it.	[492]
Practically, end users are less likely to trust and cede control to machines whose workings they do not understand [...]. [...] Investigators within the Explainable Artificial Intelligence [...] research program intend to ward off these consequences [...].	[503]
However, in order for humans to [...] trust [...] the emerging AI systems, an AI needs to be able to explain its decisions and conclusions.	[505]
Thus, it is important to build more explainable AI, so that humans can [...] effectively manage the emerging AI systems [...].	[505]
Often knowing the reasons why a particular decision has been taken [through explanations ...] can engender trust in the process that led to it [...].	[506]
Other studies [...] consistently showed that explanations significantly increase users' confidence and trust [...].	[508]
Explainable ML aims to [...] enable human users to understand, appropriately trust, and effectively manage the ML-based solutions [...].	[508]
However, too much explanation information on algorithms eroded user trust.	[508]
[...] explanation[s ...] are presented to end users [...] to boost user trust [...].	[510]
[...] participants had significantly higher trust in predictions when influences [...] were presented than those without influence information presentation [...].	[510]
Kizilcec [...] proposed that the transparency of algorithm interfaces can [...] foster user trust.	[510]
Other studies [...] consistently showed that explanations significantly increase users' confidence and trust [...].	[510]
However, too much explanation information on algorithms eroded user trust.	[510]
[...] explanations serve to build understanding and possibly trust between the AI and the user or beneficiary of the AI.	[511]
It is generally agreed upon that the goal of XAI is to increase users' trust [...].	[511]

Table 34: Quotes for the desideratum *Trustworthiness*.

Quote	Src.
[...] eXplainable Artificial Intelligence (XAI) emerged with the aims of fostering [...] trustworthiness.	[30]
Several authors agree upon the search for trustworthiness as the primary aim of an explainable AI model [...].	[54]
One way to know is to learn why the device is trustworthy by inspecting its inner workings.	[130]
[...] trustworthiness [... are] likely benefits of explanation facilities for recommender systems.	[137]
[...] the possible objectives of explanations are manifold, including aims such as increasing trustworthiness [...].	[180]
To enable [...] trustworthy integration of such systems, the end users require them to support interpretability [...] in decision-making [...].	[187]
Humans are motivated to "understand the goals, intent, contextual awareness, task limitations, [and] analytical underpinnings of the system in an attempt to verify its trustworthiness" [...].	[231]
Transparency may promote or erode users' trust in a system by changing beliefs about its trustworthiness.	[274]
Explanations enable understanding and thereby foster [...] trustworthiness [...].	[277]
Explanations help the user better understand and interpret the rationale of the recommender system, thereby making it more trustworthy and engaging.	[326]
Explanations help the user better understand and interpret the rationale of the recommender system, thereby making it more trustworthy and engaging.	[341]
Our experiments demonstrated that explanations are useful for a variety of models [...]: [...] improving untrustworthy models [...].	[393]

Continued on next page ...

Table 34 – continued from previous page

Quote	Src.
It should be clear that explicability is considered to be an important part of achieving so-called [...] 'trustworthy' [...] AI.	[397]

Table 35: Quotes for the desideratum *Understandability*.

Quote	Src.
Specific concerns that require explanations include [...] understanding system behavior.	[1]
[...] people should be able to understand how the technology may affect them [...]. [...] prior work has identified issues [...] when this is not the case. [...] To address these problems, machine learning algorithms need to be able to explain how they arrive at their decisions.	[1]
"XAI aims to ""[...] enable human users to understand [...] the emerging generation of artificially intelligent partners"."	[4]
These explanations would [...] incite the user to understand the capabilities and the limits of the agents [...].	[30]
[...] intent communication is one of the main drives for explanations in order to make the robot's internal state (e.g. goals & intentions) understandable to humans.	[30]
Explainability [...] may ease the task of elucidating the boundaries that might affect a model, allowing for a better understanding and implementation.	[54]
[...] explanation and interpretation are required to enable domain experts and users to understand [...] the [...] model [...].	[43]
[...] explainability – which is necessary for comprehensive human computing – targeting and enabling the human-in-the-loop.	[43]
[Explainable Artificial Intelligence] aims at producing intelligent systems that reinforce the trust of the users [...], who desire to understand automatic decision [...].	[48]
Explanations can serve a multiplicity of aims, including transparency (helping users to understand how the system works) [...].	[51]
Explanations can help a user understand the system's reasoning [...].	[71]
[...] work within the realm of explainability aims to help humans understand the elements of a plan suggested by the system [...].	[77]
It has been suggested that the intelligibility of system behavior is an important factor in ensuring that the user understands how the CDSS operates [...].	[93]
Previous work has shown that providing explanations can increase users' understanding of how the system operates [...]	[93]
An explanation of a decision intended to help the user understand the AI system [...] could best rely on better-world counterfactuals [...].	[95]
"To address these problems, growing work in "Explainable AI"" aims to make opaque algorithms more understandable."	[96]
Algorithmic explanations have been found to improve user understanding of the system [...].	[96]
In other words, causability is the property of the human to understand the system explanations [...].	[99]
[Explanations] may also help users to understand better all available features [...].	[108]
Explainability was shown to be a way of achieving informativeness [...]. By considering this NFR, [...] interpretability can be provided, facilitating the understanding of the system or the situation presented.	[108]
[...] explanations can be an advantage, facilitating the understanding of a system [...].	[108]
Results have shown that explanations can [...] hinder understanding if they are not displayed in a language appropriate to the specific needs of the user.	[108]
[...] explanations may add more obscurity to the understanding of the information, instead of helping to mitigate it.	[108]
[...] explanations [...] can improve users' comprehension.	[113]
[...] XAI will be key for both expert and non-expert users to enable them to have a deeper understanding [...].	[116]
Providing an explanation [...] results in better user understanding [...].	[123]
Transparency aims to increase understanding and entails offering the user insight in how a system works, for example by offering explanations for system behaviour.	[124]
Transparency can aid in the understanding of an automated vehicle's function [...].	[127]
We have argued for explainable software analytics to facilitate human understanding of machine prediction [...].	[131]
Critically, explanations are not just for people to understand the ML system [...].	[150]

Continued on next page ...

Table 35 – continued from previous page

Quote	Src.
[...] global explanations seem to render more confidence in understanding the model and generally enhance the fairness perception.	[150]
[...] using global explanations to understand and evaluate the model [...].	[150]
[...] the interfaces with explanations have a positive effect on understandability [...].	[151]
We find that including certain transparency features [...] does improve [...] understandability [...].	[161]
[...] previous research has established that explanations help users to understand [...] a recommendation [...].	[165]
Explainability [...] can help build [...] understanding [...].	[167]
[...] explanations help the human collaborator understand the circumstances that led to the behavior [...].	[167]
[...] participants preferred rationales with transparency so that they can understand [...] the robot in a situation where expectations are violated.	[167]
Explanations help the human operator understand why an agent failed to achieve a goal [...].	[166]
[...] explicability is a critical tool to build [...] understanding of [...] the technology.	[175]
[...] it is important to understand why a wrong or different decision was made: this is transparency.	[177]
[...] explanations [...] can help users to better understand the system's output [...].	[180]
[...] the explanation autonomously acquired by the IBE enriched people's understandings of the agent's future behavior [...].	[182]
[...] appropriate solutions will require an understanding [...]. This highlights the need for human-intelligible explanations of algorithmic decision making.	[193]
Besides, interaction with visualizations can strongly infuence users' understanding of complex data [...].	[203]
Thus, explanations in terms of beliefs and goals are expected to enhance trainees' understanding of the training situations.	[209]
Explanations of visual systems could also aid in understanding network mistakes [...].	[219]
[...] explanations [...] can play a key role to allow its users to better understand its outputs [...].	[221]
The explanations can be useful [...] to allow him to understand how to modify its parameters if it does not behave as expected.	[221]
The result [from adding explanations] will be filtering systems that are [...] more understandable [...].	[222]
Some of the benefits provided [by explanation facilities] are: [...] User understanding of the reasoning behind a recommendation [...].	[222]
Explanations will help users understand the process of ACF, and know where its strengths and weaknesses are.	[222]
Some of the benefits provided [by explanation facilities] are: [...] Education of the user as to the processes used in generating a recommendation, so that he may better understand the strengths and limitations of the system.	[222]
[...] self-explanation improves [...] understanding.	[231]
[...] interactive explanations will be key for understanding models better [...].	[233]
[...] data scientists use interpretability to understand the feature importance of a dataset.	[233]
[...] explainable artificial intelligence (AI) has emerged [...] to create and evaluate effective explanations for model decisions to better understand what a model has learned [...].	[233]
[...] explainability is seen as a toolset to understand the underlying technicalities and models [...].	[234]
The presence of explanations assisted participants in the conception of an accurate mental model, increasing [...] high-level understanding [...].	[235]
The benefits we see for including explanations is more information and knowledge for the user, thus [...] improving user understanding [...].	[254]
[...] transparency in design [...] may foster a better understanding of the system [...].	[274]
[...] explanations [...] help users understand its reasoning [...].	[274]
[...] explainability is concerned with enabling human understanding of various aspects of software-driven systems.	[277]
Explanations enable understanding [...].	[277]
Participants had more difficulty understanding the agent's reasoning process than the features it used, but abstract explanations of the model intelligibility type helped overcome this obstacle.	[282]
Transparency, which refers to the understandability of a specific model [...].	[303]
Our method generates contrastive explanations [...] and aims to help users understand [...] what contributed to the large error [...].	[312]

Continued on next page ...

Table 35 – continued from previous page

Quote	Src.
Explainable Machine Learning (XAI) [...] enables human users to understand [...] the emerging generation of artificially intelligent partners.	[322]
With the power of XAI or IML, new machine-learning systems will [...] provide an understanding of their future behaviour.	[322]
Explanations have an important role to play in helping users to understand the suggestions made by recommender systems.	[325]
Explanations help the user better understand and interpret the rationale of the recommender system [...].	[326]
The running hypothesis is that by building more [...] explainable systems, users will be better equipped to understand [...] intelligent agents [...].	[339]
[...] people look for explanations to improve their understanding of someone or something [...].	[339]
In many cases, an explanation [...] will [...] create a shared understanding of the decision that was made between itself and a human observer [...].	[339]
The provided explanations in these studies [...] make the recommended decision understandable.	[358]
Explaining decisions returned by intelligent systems is [...] helpful for understanding their reasoning process [...].	[372]
Explanations that help users understand how a system works have demonstrated a positive relationship with user satisfaction with the system [...].	[388]
Recently, there has been some work by the deep learning community on generating explanations as a way to better understand [...] the decisions made by deep neural networks.	[389]
Explanation can be helpful in [...] understanding and interpreting systems' output [...].	[389]
DNN developers are interested in explanation methods that allow them to understand the behavior of the DNN [...].	[391]
[...] an agent that provides an explanation for its decision might further human understanding of a medical phenomenon.	[394]
Interpretability is sometines beneficial to instill feelings of trust and understanding within the system's users.	[394]
The ability to help people understand their decisions through explanations or other means accessible to nonexperts will provide people with greater sense of trust [...].	[395]
[...] explanations were needed to help understand a system malfunction [...].	[399]
Explaining the reason for recommendations [...] helps a user understand why an item is recommended.	[410]
[...] explanations show a significant improvement in [...] understanding [...].	[412]
[...] more explanation facilities appeared to help participants understand [...] Clarisense's search strategy.	[414]
[...] insights into the inner workings of a trained model allow users and analysts, [...] to understand the models [...].	[422]
An intelligent robot that is explainable yields several important advantages. [...] systems that [users] understand – or at least believe that they understand.	[426]
[...] explanations [...] help a user to understand the history and experience behind the decision.	[425]
The role of explanation is to help the user to better understand the agent's behaviour [...].	[427]
Interpretability of the machine learning predictions is important for a variety of reasons. [...] the users have to [...] understand the underlying decisive mechanisms [...].	[441]
The goal of an explanation of this kind is to impart an understanding of how the system found an answer.	[444]
Such techniques [of explanation] can thus be essential contributors to [...] providing [...] information to keep the agents on the same page with respect to their understanding of each others' tasks and capabilities [...].	[451]
There is considerable work on making the decisions of an existing learned model or reasoning system more interpretable [...] to make decisions that are easier for humans to understand.	[452]
[...] computer-generated visualizations can increase our understanding of the models they depict.	[454]
Explanations should be part of a cycle, where the user understands what is going on in the system [...].	[465]
[...] goals in the ML domain, namely understanding, [... are] related to the problems of interpretability of the ML results and comprehensibility of the obtained models.	[476]
Users who are using the system to accomplish personal end goals [...] will likely seek information that assists their understanding of how the system processes data to arrive at its outputs [...].	[482]
[...] explanations can improve trust because of increased user understanding [...].	[485]
To enable end users to understand [...] their intelligent partners, [...] researchers have produced many [...] algorithm visualizations, interfaces and toolkits [...].	[485]
[...] they initially found that Why and Why Not explanations were most effective in promoting system understanding [...].	[485]

Continued on next page ...

Table 35 – continued from previous page

Quote	Src.
We list several types and goals of transparency. [...] For a developer, to understand how their system is working [...].	[490]
We list several types and goals of transparency. [...] For society broadly to understand and become comfortable with the strengths and limitations of the system [...].	[490]
We list several types and goals of transparency. [...] For a user to understand why one particular prediction or decision was reached [...].	[490]
Human interpretability – that is helping humans to understand machines – is of great importance	[490]
[...] an explanation for an end-user is intended [...] to aid the user in understanding the consequences of the system's conclusion.	[493]
For an end-user audience, this purpose is to help the end-user better understand the domain in which the expert system is operating	[493]
However, in order for humans to understand [...] the emerging AI systems, an AI needs to be able to explain its decisions and conclusions.	[505]
Thus, it is important to build more explainable AI, so that humans can [...] effectively manage the emerging AI systems [...].	[505]
Explainable ML aims to [...] enable human users to understand, appropriately trust, and effectively manage the ML-based solutions [...].	[508]
[...] explanation [...] was investigated to allow users to understand why a classification/prediction is made.	[510]
[...] explanations serve to build understanding [...] between the AI and the user or beneficiary of the AI.	[511]

Table 36: Quotes for the desideratum *Usability*.

Quote	Src.
Specific concerns that require explanations include usability [...].	[1]
[...] explanations can serve a multiplicity of aims, such as [...] increasing the ease of use of a system [...].	[51]
The ability of recommender systems to effectively explain their recommendations is a potentially crucial aspect of their [...] usability.	[70]
Explanations may also help to improve the usability of the system, easing the use and teaching the user how to better operate it.	[108]
[...] explanations can be a good way to mitigate the complexity of the system and help the user to better operate it.	[108]
[...] usability [...] of a personalized mobile application depends highly on perceived information transparency [...].	[112]
Explanation [...] can substantially affect [...] usability [...].	[137]
Interpretability is used to confirm other important desiderata of ML systems: [...] Usable methods provide information that assist users to accomplish a task.	[155]
[...] we argue that interpretability can assist in qualitatively ascertaining whether other desiderata such as fairness, privacy, reliability, robustness, causality, usability and trust are met.	[155]
[...] ease-of-use [... was] strongly affected by the presence of justification-type explanations [...].	[199]
[...] subjects receiving justification explanations rated the system significantly easier to use [...].	[199]
Another property [...] is usability: people tend to trust more models providing information that rassist them to acocmplish a task [...].	[201]
[...] explanations must present easy-to understand coherent stories in order to ensure good use of the AI [...].	[231]
[...] explainable-AI [...] brings usability [...] into a new and important focus [...].	[236]
The benefits we see for including explanations is more information and knowledge for the user, thus hopefully easing the interaction [...].	[254]
Explanations enable understanding and thereby [...] improve usability [...].	[277]
The third theme of motivation for explainability is to adapt usage or interaction behaviors to better utilize the AI.	[304]
Explanation Purposes Identified in Primary Studies [...]: [...] Increase the ease of use [...].	[358]
Explanations improve system usability [...].	[388]
The role of explanation is to help the user to better understand the agent's behaviour, [...] to make better use of it by performing "model reconciliation" [...].	[427]

Continued on next page ...

Table 36 – continued from previous page

Quote	Src.
Among other things, good explanations could [...] make it quicker and easier for users to find what they want [...].	[463]
Possible aims for explanations: [...] Increase the ease of usability or enjoyment.	[464]
Among other things, good explanations could [...] make it quicker and easier for users to find what they want [...].	[464]
Explanations can also serve other aims such as [...] make it quicker and easier for users to find what they want [...].	[465]
Explanatory criteria and their definitions: Satisfaction [...]. Increase the ease of use or enjoyment.	[465]
The presence of longer descriptions of individual items has been found to be positively correlated with [...] ease of use of the recommender system [...].	[465]
Explanatory aims: [...] Increase the ease of use [...].	[466]
[...] on-demand explanation [...] increases [users'] perception of system explainability. However, the improvement comes with a price of reducing [...] the sense of ease of use [...].	[473]
[...] the goal of XAID includes investigating the actual usability of XAI in terms of how it supports game designers in specific design tasks.	[511]

Table 37: Quotes for the desideratum *Usefulness*.

Quote	Src.
The ability of recommender systems to effectively explain their recommendations is a potentially crucial aspect of their utility [...].	[70]
[...] explanations can significantly increase [...] perceived information usefulness [...].	[111]
[..] studies [...] may focus on how explanations enhance perceived usefulness [...].	[165]
[...] explanations [...] promote objectives such as [...] utility.	[180]
[...] good explanations can [...] make it easier for users to find what they want.	[198]
[...] perceptions of usefulness [...] were [...] strongly affected by the presence of justification-type explanations [...].	[199]
[...] explanations [...] can play a key role [...] to allow its users to better understand its outputs and therefore to make a better use of it.	[221]
The system [...] provides knowledge and explanations necessary for the user to carry out his or her task [...].	[254]
[...] participants felt that [an organization-based explanation interface] would be easier for them to compare different products and make a quicker decision.	[382]
How explanations can also increase [...] the perceived usefulness of a system [...].	[388]
The expected impacts of [...] explanations are as follows: [...] usefulness: envisioning a context helps users to make the right choices [...].	[410]
We further confirmed that the hybrids of the context style and other explanation styles improve persuasiveness and usefulness.	[410]
The presence of longer descriptions of individual items has been found to be positively correlated with both the perceived usefulness [...].	[465]
[...] knowledgeable explanations significantly increase the perceived usefulness of a recommender system [...].	[502]
The findings show that the explanation feature can significantly increase a recommender system's perceived usefulness [...].	[502]

Table 38: Quotes for the desideratum *Validation*.

Quote	Src.
Explanations [...] provide a way to verify the validity of a decision.	[71]
[...] explanations are of uttermost importance [...] to ensure that algorithms are performing as expected [...].	[99]
[Explanation facilities] were perceived as being better suited to knowledge engineers - for validating system knowledge [...].	[137]
Explanations can also be used to ascertain whether certain criteria were used appropriately or inappropriately in case of a dispute.	[156]

Continued on next page ...

Table 38 – continued from previous page

Quote	Src.
While some might wholesale reject the schema of classifications used, others might want to know if such a decision was made soundly. For these decision subjects, an explanation might help.	[164]
Numerous studies have demonstrated the benefits of explanations in intelligent systems [...], including [...] supporting the evaluation of system conclusions [...].	[165]
These explanations are important [...] to ensure that the algorithms perform as expected.	[188]
Questions that experts in artificial intelligence (AI) ask opaque systems provide inside explanations, focused on [...] validation.	[189]
These outside explanations can [...] act as external validation	[189]
[...] classification systems may generate explanations in order to [...] satisfy the user of its validity.	[325]
The main goal of Explainable Artificial Intelligence (XAI) has been variously described as [...] validating the decision process of an opaque AI system [...].	[365]
[...] another reason is quality, which an explanation can help validate.	[489]

Table 39: Quotes for the desideratum *Verification*.

Quote	Src.
Explanations [...] provide a way to verify [...] a decision.	[71]
[...] study participants wanted better explanations to help them [...] verify that the disorder fit the suggestion [...].	[93]
"[...] “how” explanations [... enable] users to verify that an algorithm has accurately and fairly ""produc[ed] and certif[ied] knowledge""."	[123]
[...] system developer [...] may use an expert system explanation facility during the development phase to verify the correctness of the knowledge base [...].	[137]
[...] studies applying the transparency as verifiability approach tend to find positive outcomes for organizations.	[173]
Revealing the reasoning makes it possible for stakeholders to spot possible flaws and also to identify whether the line of reasoning results in outcomes that match the disclosed data.	[241]
Our work also showcases the fact that interpretability is [...] a powerful tool [...] for verifying predictions [...].	[322]
[...] explanations are used to evaluate the proposed solution and to justify changes to this solution.	[325]
For all types of algorithms, auditing is a necessary precondition to verify correct functioning.	[342]
Explanations can be necessary to [...] verify and improve the functionality of a system [...].	[341]
The question of what kinds of explanation a human can utilize implies the presence of a downstream task. [...] Intrinsic tasks include goals such as verification – given an input, output, and explanation, can the human user verify that the output is consistent with the input and provided explanation?	[350]
The [...] purpose of explanations [...] to help users assess if the recommended alternative is truly adequate for them.	[358]
[...] the use of AI techniques that do not readily provide these [explanation] capabilities should be used with extreme caution as verification of their performance is going to be at least in part dependent on statistical behaviour.	[427]
Experts tend to require explanations to verify the reasoning of the system and explain away surprising results.	[444]
They suggest three major explanation goals. Verification is the goal of the knowledge engineer in verifying that the system works as it should.	[444]
The goal of an explanation [...] is to impart an understanding of how the system found an answer. This allows the users to check the system by examining the way it reasons [...].	[444]
We list several types and goals of transparency. [...] For a user [...] to allow a check that the system worked appropriately [...].	[490]
There are three major explanation goals, namely verification [...]. Within the context of verification, the goal [...] is to verify the knowledge of the expert system.	[493]

D. Deutsche Zusammenfassung

Einleitung

Ob Pflegeroboter in Japan, selbstfahrende Busse in Deutschland oder automatisierte Personalauswahlsysteme in den Vereinigten Staaten von Amerika—komplexe, künstliche Rechnersysteme[84] sind aus unserem Alltag nicht mehr wegzudenken. Aus dieser Entwicklung ergeben sich zwei große Herausforderungen: *Maschinenethik* und *Maschinenerklärbarkeit*. Maschinenethik befasst sich mit Verhaltensbeschränkungen für solche Systeme, um ein eingeschränktes, moralisch akzeptables Verhalten zu gewährleisten; Maschinenerklärbarkeit sucht nach Möglichkeiten, die Handlungen und Entscheidungsprozesse von Systemen zufriedenstellend zu erklären, so dass ihre menschlichen Benutzer diese Systeme verstehen und sich ihrer gesellschaftlich nützlichen Auswirkungen sicher sein können.

Maschinenethik und Maschinenerklärbarkeit erweisen sich nur in Symbiose als besonders effizient. Vor diesem Hintergrund zeigen wir in dieser Arbeit, wie Maschinenethik Maschinenerklärbarkeit voraussetzt und wie Maschinenerklärbarkeit Maschinenethik einschließt. Wir entwickeln diese beiden Facetten anhand der oben genannten Beispiele. Anhand dieser Beispiele argumentieren wir für eine spezifische Sichtweise der Maschinenethik, die sogenannte *moralische Ausrichtung*, und schlagen vor, wie diese in einem Framework formalisiert werden kann. Im Hinblick auf Maschinenerklärbarkeit legen wir dar, wie das von uns vorgeschlagene Framework durch die Verwendung eines argumentationsbasierten Ansatzes für die Entscheidungsfindung eine Grundlage für Maschinenerklärbarkeit bieten kann.

Obwohl einige Forscher glauben, dass fest verdrahtete moralische Einschränkungen in Bezug auf das Verhalten von Maschinen eine ausreichende Voraussetzung dafür sind, dass Menschen vernünftiges Vertrauen in künstliche Systeme entwickeln können, möchten wir erörtern, warum dies nicht der Fall ist. Stattdessen halten wir es für notwendig, die Maschinenethik durch Mittel zu ergänzen, mit denen wir feststellen können, ob das Vertrauen, das wir in solche Systeme setzen, gerechtfertigt ist und ob sie andere wünschenswerte Eigenschaften haben. Nachdem wir dargelegt haben, warum dies wichtig ist, argumentieren wir, dass es mindestens eine passende Ergänzung für die Maschinenethik zu diesem Zweck gibt: Maschinenerklärbarkeit—die Entwicklung von Mitteln, mit denen die Handlungen und Entscheidungsprozesse künstlicher Systeme erklärt werden können.

Maschinenerklärbarkeit trägt also zur Maschinenethik bei. Diese Beziehung besteht auch umgekehrt: Maschinenethik trägt zu Maschinenerklärbarkeit bei, da Maschinenerklärbarkeit besonders gut mit einem moralischen System als Grundlage für die Erzeugung von Erklärungen gedeihen kann. Eingebettet in ein moralisches System können Erklärungen auf moralische Erwägungen Bezug nehmen und so einen hervorragenden Ausgangspunkt für Vertrauen (und andere wünschenswerte Eigenschaften) bieten.

[84] Im Rahmen dieser Zusammenfassung benutzen wir die Begriffe “Künstliches Rechnersystem”, “künstliches System” und “Maschine” synonym (genauso wie die entsprechenden Englischen Begriffe in der Arbeit).

Das übergeordnete Ziel dieser Arbeit ist es, genau eine solche Verbindung zwischen Maschinenethik und Maschinenerklärbarkeit herauszuarbeiten. Es gibt jedoch auch andere Ziele, die wir untersuchen. Die Arbeit ist zu diesem Zweck in vier Teile gegliedert, von denen die ersten drei weitgehend unabhängig voneinander gelesen werden können. Jeder Teil trägt nicht nur zum übergeordneten Ziel dieser Arbeit bei, sondern hat auch eigene, untergeordnete Ziele. Im Folgenden wollen wir die vier Teile der Arbeit kurz zusammenfassen, um einen Überblick über die Arbeit zu gewährleisten. Es sollte angemerkt sein, dass diese Zusammenfassung die Gesamtarbeit natürlich nur unzureichend widerspiegeln kann.

Maschinenethik

Der erste Teil dieser Arbeit beschäftigt sich mit Maschinenethik. Maschinenethik hat sich in den letzten Jahren zu einem ernstzunehmenden Forschungsgebiet entwickelt, zu dem mittlerweile auch die ersten systematischen Arbeiten veröffentlicht wurden (z.B. [22, 484]). Insgesamt ist der genaue Forschungsgegenstand der Maschinenethik jedoch eine offene Frage.

Vor diesem Hintergrund wies James H. Moor darauf hin, dass der Begriff “Maschinenethik” recht weit gefasst werden kann. Ihm zufolge reicht das Verständnis von der Implementierung moralisch motivierter Beschränkungen für das Verhalten komplexer und möglicherweise autonomer künstlicher Systeme bis hin zur Implementierung vollwertiger moralischer Fähigkeiten [346]. Einerseits ist die erste Sichtweise bereits heute von großer praktischer Bedeutung, da der moralische Einfluss, der durch künstliche Systeme sowohl direkt als auch indirekt ausgeübt wird, stetig zunimmt. Andererseits befasst sich die letztgenannte Sichtweise mit Szenarien, die Science–Fiction bleiben, und beinhaltet Diskussionen über tiefgreifende philosophische Konzepte von Autonomie und freiem Willen.

Der erste Teil dieser Arbeit setzt sich mit Maschinenethik auseinander. Insbesondere untersuchen wir, ob Maschinenethik ein lohnendes Unterfangen ist. Dabei argumentieren wir für eine bestimmte Sichtweise der Maschinenethik: *moralische Ausrichtung*. Darüber hinaus untersuchen wir auch die Verbindung zwischen Maschinenethik und Maschinenerklärbarkeit genauer. Schließlich erörtern wir, wie eine sinnvolle Implementierung von Maschinenethik aussehen sollte. Im Einzelnen sieht dieser Teil der Arbeit wie folgt aus.

In Kapitel 2 stellen wir Maschinenethik anhand eines allgemeinen Überblicks über das Forschungsgebiet der Ethik als einen Zweig der angewandten Ethik vor. Zudem geben wir einen Überblick über die beiden Hauptforschungsströmungen der Maschinenethik. Den Abschluss des Kapitels bildet eine kurze Vorstellung einiger Ansätze, die in den letzten Jahren im Bereich der Maschinenethik entwickelt worden sind.

Darauf aufbauend untersuchen wir Maschinenethik in Kapitel 3 genauer. Insbesondere diskutieren wir, ob es sich lohnt, zu diesem Thema zu forschen, oder ob eine solche Forschung mehr Nachteile als Vorteile bringen könnte. Indem wir unseren eigenen Standpunkt zur Maschinenethik darlegen, zeigen wir, dass sich diese Forschung in der Tat lohnt, vor allem, wenn sie durch Forschung zur Erklärbarkeit von Maschinen unterstützt wird.

Kapitel 4 bildet den Abschluss unserer Diskussion zur Maschinenethik. Hier schlagen wir eine Brücke zu dem von uns im zweiten Hauptteil der Arbeit angestrebten Framework, indem wir erste Überlegungen dazu anstellen, wie ein solches Framework aussehen sollte. Konkret plädieren wir für einen prinzipienbasierten Ansatz und liefern Argumente für einen solchen. Darüber enthält dieses Kapitel direkte Forschung zur Maschinenethik, indem es eine Diskussion über die Vorteile und Nachteile der Programmierung eines künstlichen Systems mit einer der traditionellen normativen Theorien beinhaltet.

Formale Maschinenethik

Im zweiten Teil dieser Arbeit bauen wir auf den vorherigen Argumenten auf und skizzieren ein entsprechendes prinzipienbasiertes Framework für Maschinenethik. Unser Framework ist im mathematischen Sinne *formal*, indem es eine Sammlung systematisch ausgearbeiteter Ideen und Strukturen bietet, die es ermöglichen, ein künstliches System mit einem *instrumentellen* Ziel und *normativen* Einschränkungen zu beschreiben. Darüber hinaus ist unser Framework *generell*, in dem Sinne, dass wir versuchen, es unabhängig von den Annahmen bestimmter theoretischer Perspektiven der normativen Ethik zu motivieren.

Wir denken, dass es von entscheidender Bedeutung ist, dass wir keine spezifischen normativen Einschränkungen vorschlagen, um so wenige Einschränkungen wie möglich zu haben. Unser Framework ist dadurch flexibel genug, um später mit spezifischen normativen Einschränkungen ausgefüllt zu werden. Dieser Ansatz ist unter anderem dadurch motiviert, dass wir der Meinung sind, dass die Frage nach der Angemessenheit von normativen Einschränkungen von der Domäne abhängen kann, in der ein System eingesetzt werden soll.

Über das formale Framework hinaus, aber auf diesem aufbauend, schlagen wir einen argumentationsbasierten Ansatz zur Entscheidungsfindung von künstlichen Systemen vor. Dieser Ansatz hat den Vorteil, dass er zur Generierung von Erklärungen verwendet werden kann. Indem das Framework dadurch sowohl Maschinenethik gewährleistet als auch Maschinenerklärbarkeit ermöglicht, bildet es einen wichtigen Baustein für unser Argument, dass die beiden Bereiche eng miteinander verbunden sind. Im Einzelnen sieht Teil 2 wie folgt aus.

Wir beginnen den zweiten Teil dieser Arbeit in Kapitel 5 mit der Entwicklung unseres Frameworks für Maschinenethik. Dieses Framework basiert im Kern auf frei nach Anwendungsfall wählbaren moralischen Prinzipien, die eine wesentliche Rolle bei der Entscheidungsfindung spielen sollen. Zudem enthält das Framework die Möglichkeit, die instrumentellen Ziele eines Systems bei der Entscheidungsfindung unter Unsicherheit zu berücksichtigen. Dadurch soll ein normativ eingeschränktes System erreicht werden, welches nützlich bleibt.

Auf diesem Framework aufbauend beschreiben wir in Kapitel 6 einen detaillierte Entscheidungsfindungsprozess. Dieser Prozess stützt sich auf Argumente in einem graphenbasierten Ansatz, der eine fruchtbare Grundlage für die Erstellung von Erklärungen bietet. Ziel ist es dabei, die instrumentelle und normative Eignung des Systems gleichzeitig zu optimieren.

Den Abschluss des zweiten Teils der Arbeit bildet Kapitel 7 mit einigen zusätzlichen Gedanken und Ideen zum Framework. Insbesondere diskutieren wir Ansätze zur Modellierung der moralischen Prinzipien, auf denen das Framework aufgebaut ist. Zudem skizzieren wir eine alternative, auf STIT-Logik aufbauende Formalisierung für das Framework.

Maschinenerklärbarkeit

Über Maschinenethik hinausgehend ergibt sich die Notwendigkeit von Maschinenerklärbarkeit. Insbesondere im Zusammenhang mit künstlichen Systemen (die oft positive gesellschaftliche Auswirkungen versprechen) sollte man Black-Box-Systemen, deren Entscheidungen, Vorhersagen oder Verhalten wir nicht genau erklären können, auf Dauer nicht vertrauen. Viele Anwendungen künstlicher Systeme—zum Beispiel als Berater von Politikern und Richtern—setzen mehr als undurchsichtige Ergebnisse wie Zahlen (und insbesondere Wahrscheinlichkeiten) voraus, zumindest im Kontext liberaler Demokratien. Diese Systeme müssen überprüfbar sein, und ihre Ergebnisse müssen zumindest prinzipiell und auf Anfrage begründbar sein.

Selbst unter der Prämisse, dass der Einsatz einiger künstlicher Systeme aus moralischer Sicht wünschenswert ist (z.B. wegen ihrer Gesamtwirkung), und selbst wenn sich diese Systeme tatsächlich so moralisch gut verhalten würden, wie es logisch und konzeptionell möglich ist, solange die Menschen diesen Systemen nicht gerechtfertigt vertrauen können und keinen Zugang zu den Gründen für ihre Entscheidungen haben, ist der Einsatz dieser Systeme selbst dort gefährdet, wo er wünschenswert wäre, und kann in vielen potenziell vielversprechenden Anwendungsbereichen nicht mit gutem Gewissen gefördert werden.

Der dritte Teil der Arbeit beschäftigt sich daher mit Maschinenerklärbarkeit. Da die Forschung zur Maschinenerklärbarkeit noch sehr jung ist, geht es in diesem Teil der Arbeit vor allem um Klarstellungen. Insbesondere extrahieren wir die Ziele der Maschinenerklärbarkeit aus der Literatur und fassen sie in einem Modell zusammen. Diese Ziele sind wiederum mit der Maschinenethik verknüpft und bilden den dritten Bestandteil unserer Argumentation für die enge Verbindung zwischen Maschinenethik und Maschinenerklärbarkeit. Zum Abschluss der Darstellung der Maschinenerklärbarkeit stellen wir einige Ansätze vor, die auf Erklärbarkeit abzielen, und entwickeln Qualitätskriterien für diese Ansätze.

In Kapitel 8 geben wir einen Überblick über die Forschung zu Erklärungen im Allgemeinen, um diese zur Forschung zu Maschinenerklärbarkeit abzugrenzen. Das Konzept der Erklärung im Bereich der Maschinenerklärbarkeit scheint ein pragmatisches zu sein. Basierend auf dieser Erkenntnis stellen wir unser Modell der Hauptprozesse in der Maschinenerklärbarkeit vor und zeigen, wie sie miteinander in Beziehung stehen. Unsere Vorstellung ist es, dass Ansätze zur Förderung der Erklärbarkeit im Einklang mit dem pragmatischen Begriff der Erklärungen im Bereich der Maschinenerklärbarkeit darauf abzielen, das Verstehen von verschiedenen Adressaten zu erhöhen. Dies ist jedoch nicht das übergeordnete Ziel solcher Ansätze. Wir gehen davon aus, dass Erklärbarkeitsansätze letztlich darauf abzielen, bestimmte übergeordnete Ziele zu erreichen, wie z.B. Fairness und gerechtfertigtes Vertrauen.

Zur weiteren Vertiefung des Modells kommen wir in Kapitel 9 auf die Verbindungen zur Maschinenethik zurück. Anhand einer umfangreichen systematischen Literaturanalyse von mehr als 200 Publikationen zeigen wir die Ziele der Forschung im Bereich der Maschinenerklärbarkeit auf, von denen viele eine moralische Komponente haben. Bei dem Ziel, Verantwortung für Entscheidungen, die auf den Empfehlungen von künstlichen Systemen beruhen, adäquat zuschreiben zu können, gehen wir näher darauf ein, wie Maschinenerklärbarkeit zu ihm beitragen soll. Dazu gucken wir uns einen spezifischen Fall an, in dem ein Personaler mithilfe eines Vorauswahlsystems eine Person einstellen soll.

Den Abschluss der Diskussion zur Maschinenerklärbarkeit bildet Kapitel 10, in dem wir Ansätze vorstellen, die künstliche Systeme erklären sollen. Nachdem wir einige exemplarische Ansätze vorgestellt haben, wenden wir uns der Bewertung solcher Ansätze zu. Dazu entwickeln und motivieren wir unsere eigenen Qualitätskriterien für Erklärungen. Diese Kriterien sind Treue (d.h. die Erklärung sollte sich auf die realen Gründe für die Entscheidung eines Systems beziehen), Verständlichkeit (d.h. die Erklärung sollte verständlich sein), und Bewertbarkeit (d.h. die Erklärung sollte es ermöglichen, zu überprüfen, ob das System bestimmte Kriterien, wie beispielsweise Fairness, erfüllt). Als Argument für diese Qualitätskriterien führen wir an, dass sie besonders gut dazu geeignet sind, die übergeordneten Ziele der Maschinenerklärbarkeit zu erreichen. Abschließend wenden wie die Kriterien auf die diskutierten Ansätze an und stellen fest, dass die meisten Ansätze die Kriterien nicht erfüllen.

Abschließender Brückenschlag

Der vierte und letzte Teil der Arbeit verbindet die ersten drei Teile miteinander. Unter Berücksichtigung unserer Ansichten über Maschinenethik und Maschinenerklärbarkeit wenden wir die Qualitätskriterien, die wir für Erklärbarkeitsansätze entwickelt haben, auf die potenziell durch unser Framework generierbaren Erklärungen an. Dadurch können wir für die Angemessenheit unsere Frameworks argumentieren und seine Vergleichbarkeit mit anderen Ansätzen aufzeigen. Trivial ist diese Argumentation jedoch nicht, da die Erklärungen, die mithilfe des Frameworks generiert werden können, höchstwahrscheinlich sehr formal sein werden und so insbesondere für Laien unbrauchbar sein dürften. Hier bildet die Forschung zu Idealisierung und Abstraktion aus der Wissenschaftstheorie einen wichtigen Argumentationsbestandteil.

Im letzten Kapitel dieser Arbeit fassen wir kurz unsere Hauptstandpunkte zusammen. Darauf folgt ein größerer Überblick über Möglichkeiten für zukünftige Forschung. Insbesondere beim Framework mussten wir viele Vereinfachungen vornehmen, die in einer zukünftigen Version des Frameworks ausgearbeitet werden sollten. Zudem mussten wir viele Annahmen machen, beispielsweise darüber, wie man Gründe gegeneinander gewichtet, die es weiter zu untersuchen und zu fundieren gilt. Schließlich hat die Effizienz unseres Entscheidungsfindungsalgorithmus noch Optimierungsbedarf, den es zukünftig durch sinnvolle Heuristiken auszuloten gilt.

E. Supplementary Information about the Thesis

Origin of the Sections In order to meet scientific standards and make transparent all sources used in this thesis, please find below a comprehensive overview of how each (sub)section has been influenced by other works of mine. Please refer to the foreword for more information on how this thesis differs from these works.

Abstract The thematic proximity alone makes it inevitable that parts of the abstract are based on the abstract of my master's thesis [446].

Section 1 This section is a revised and modified version of Section 1 of my master's thesis [446]. My master's thesis, in turn, builds on the publications highlighted in the foreword. A more detailed breakdown can be found in the foreword of my master's thesis.

Section 4 The introduction of this section builds in part on Section 1.1 of my master's thesis [446]. In addition, a few arguments in Section 4.1.2 were adapted from arguments presented in Section 3.1.2 of my master's thesis.

Section 5 This section revises, improves, and expands upon the considerations I present in Section 2 and Section 3 of my master's thesis [446]. In particular, Section 5.1 builds on parts of Section 2.1 (excluding Section 2.1.2 and Section 2.1.3), Section 5.2 builds on Section 2.2, Section 5.3 builds on Section 3.1, and Section 5.4 builds on Section 3.2.

Section 6 This section revises, improves, and expands upon the considerations I present in Section 5 of my master's thesis [446]. In particular, Section 6.1 builds on Section 5.1, Section 6.2 builds on Section 5.2, and Section 6.3 builds on Section 5.3.

Section 7 Section 7.1 builds on Section 6.1, Section 6.2, and Section 6.3 of my master's thesis [446]. Section 7.2 brings together ideas that Kevin Baum and I had for revising the AMAI paper (see the foreword) for the first time in a sufficiently mature version.

Section 8 The introduction of Section 8.1 builds on Section 4.1 of my master's thesis [446]. While Section 8.1.1 is new, Section 8.1.2 builds on Section 5 of [63]. Finally, Section 8.4 builds on smaller parts of each Section 1, Section 2, and Section 3 of [294]. In particular, Section 8.4.2 builds in larger parts on Section 2.1 of [294].

Section 9 The introduction of Section 9 and parts of Section 9.1 build on Section 2 of [294]. While the basis for Section 9.2.1 is the systematic literature review I conducted with other researchers for [105], it should be noted that I have revised and re-evaluated this literature review to a significant extent. More information on this revision can be found in Appendix C. Nevertheless, some parts of Section 6 of [105] have still been used as a basis for Section 9.2.1 of this thesis. The same is true for Section 2.2 and Section 3.3 of [294]. Finally, Section 9.3 builds on Section 4 of [63].

Section 10 Section 10.1.1 originates from Section 5 of [294]. The rest of Section 10.1 is slightly enriched with content from Section 4.4 of my master's thesis [446]. Section 10.2 builds on [448], but is enriched in Section 10.2.2 by content from Section 4.2 of my master's thesis and in Section 10.2.3 by content from Section 4.5 of my master's thesis.

Section 11 The introduction of Section 11 builds on Section 5.4 of my master's thesis [446], as does Section 11.1.1, the introduction of Section 11.2, and Section 11.2.3.

Section 12 As with the abstract, it is unavoidable that this section is a revised and updated version of Section 9 of my master's thesis [446] due to its thematic proximity.

Origin of the Figures Next, I would like to turn to the figures and show for each its origin. Figure 1 is inspired by slides from the lecture "Ethics for Nerds" by Kevin Baum and Sarah Sterz. Figure 2 was created together with Kevin Baum for [60]. Figure 3 was created together with Kevin Baum for [61]. Figure 4 was created together with Kevin Baum for the initial version of the AMAI paper. Figure 5 is taken from [332]. Figure 6, Figure 7, Figure 23, Figure 25, and Figure 26 are adapted from [171]. Figure 8 and Figure 15 are my own creation. The subfigures of Figure 9 are adapted from the sources indicated at them respectively (viz., [399], [263], and [116]). Figure 11 and Figure 12 were created together with Markus Langer, Daniel Oster, and Kevin Baum for [294]. Figure 18 is adapted from [450]. Figure 24 was created together with Daniel Oster for research purposes. Finally, Figure 32 is a screenshot of a Google Trends search, made on the 20th of November, 2022.

Figure 10, Figure 13, Figure 14, Figure 16, Figure 17, Figure 19, Figure 20, Figure 21, Figure 22, Figure 27, Figure 29 and Figure 28 were created by the Python programs listed in Appendix A (more on the origin of these programs in the paragraph below). In particular, Figure 10a, Figure 13, Figure 14, Figure 27, Figure 28, Figure 29 were created by the program described in Listing 1; Figure 10c, Figure 16b, Figure 17, and Figure 19 were created by the program described in Listing 2; Figure 16a and Figure 31 were created by the program described in Listing 3; Figure 10b, Figure 16c, Figure 20, and Figure 30 were created by the program described in Listing 4; Figure 21 was created by the program described in Listing 5; and Figure 22 was created by an unrecoverable version of the program described in Listing 6.

The underlying original images are all taken from unsplash.com. Thus they are under the unsplash license, which permits free use, even for commercial purposes. The dome image is available under https://unsplash.com/photos/kQ80v_7PjRs and courtesy by Nik Nikolla. The lion image is available under https://unsplash.com/photos/BBzTMPUSAO0 and courtesy by Joshua J. Cotten. The lynx image is available under https://unsplash.com/photos/ToP7JBTcsfY and courtesy by Zdeněk Macháček. Finally, the elephant image is courtesy by Brianna R. and available under https://unsplash.com/photos/5dsApXnqtEk.

Origins of the Code Listings The Python programs used to create the example images for the different explainability approaches can be found in Appendix A. These programs originate from various internet sources and I have adapted them for my purposes. In particular, I have adapted them to use Google's InceptionV3 model [458] as the AI model whose predictions are explained. By using the same model for all explainability approaches, the explanatory information generated by these approaches is nicely comparable.

The exact origins are as follows: Listing 1 is adapted from [36]; Listing 2 is adapted from [194]; Listing 3 is adapted from [353]; Listing 4 is adapted from [459]; Listing 5 is adapted from [491]; and Listing 6 is adapted from [87]. In each case, the licenses of the originals are preserved and can be looked up in the indicated sources.

Glossary

artificial intelligence (AI) is the set of theories and techniques used to create machines capable of simulating human intelligence (e.g., understanding language, recognizing pictures). 20, 33, 76, 93, 125, 127, 136, 140, 142, 143, 148, 152, 175, 181, 263, 265, 266, X

artificial neural network (ANN) is a type of AI that tries to imitate the human brain. 18, 19, 129, 158, 161–164, 166, 167, 171, 173, 265, 266, X

autonomous vehicle (AV) is a self-driving vehicle. 21, 24, 28, 29, X

class activation map (CAM) is an explainability approach that produces heatmaps for classifications of CNNs. 161, 164, 165, 172, 174, 201, 203, 211, V, IX, X

change impact analysis (CIA) is an explainability approach that analyzes counterfactual situations. 158, 159, 172–174, X

causal-mechanical explanation (CME) is a form of scientific explanation based on causal relationships. 116, 117, 126, 181, X

convolutional neural network (CNN) is a type of ANN specifically fit for image recognition. 161, 162, 164–167, 265, X

deep learning (DL) is a type of ML for creating DNNs. 126, 181, X

deductive-nomological explanation (DNE) is a form of scientific explanation based on laws and regularities. 116–118, 126, 181, 183, 184, 186, X

deep neural network (DNN) is a type of ANN with many (hidden) layers. 35, 38, 52, 160, 169, 265, X

feature visualization (FV) is an explainability approach that reveals what (clusters of) neurons in an ANN respond to by generating an input that fully activates them. 165, 167, 173, 174, 207, 209, V, IX, X

high level expert group on artificial intelligence (HLEGAI) is the expert group on AI appointed by the European Commission. 148, X

human resources (HR) are intangible resources that a company derives from its employees. 152, X

local interpretable model-agnostic explanation (LIME) is a well-known explainability approach for all types of models that explains individual predictions. 159–162, 165, 171–175, 193, 195, 210, V, IX, X

machine learning (ML) is a field of study in AI that relies on mathematical and statistical approaches to give computers the ability to "learn" from data, that is, to improve their performance in solving tasks without being explicitly programmed for each. 17, 19, 22, 23, 44, 45, 51–53, 93, 126, 127, 130, 136, 139, 140, 142, 145–147, 149, 150, 158, 160, 167, 169, 171, 175, 188, 190, 265, X

new mechanist explanation (NME) is a form of scientific explanation based on the interplay of mechanisms. 116, 117, 126, 181, X

standard deontic logic (SDL) is the most cited and studied system of deontic logic [332]. 101–104, X

testing with concept activation vectors (TCAV) is a peculiar explainability approach for ANNs that allows to examine the importance of so-called "concepts" on prediction classes. 165, 166, 171, 174, 175, 203, V, X

explainable artificial intelligence (XAI) is the research field concerned with making AI explainable. 127, X

moral agency is the property of being a moral agent (i.e., the property of being an entity that can act based on morals). 14, 15

moral agent is an entity that can act based on morals. 14, 15, 25, 41–43

moral patiency is the property of being a moral patient (i.e., the property of being an entity that should be included in moral considerations). 15, 16, 24

moral patient is an entity that has moral relevance and should be factored in when making moral judgments. 9, 15, 17, 24, 27, 33, 47, 51, 54

superintelligence is a hypothetical entity that possesses intelligence far surpassing that of the brightest and most gifted human minds. 22, 24, 29, 33, 34, 38

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda". In: *Proceedings of the 36th Conference on Human Factors in Computing Systems*. CHI 2018 (Montréal, Québec, Canada). Ed. by Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox. New York, NY, USA: Association for Computing Machinery, 2018, 582, pp. 1–18. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174156.

[2] Peter Achinstein. "What Is an Explanation?" In: *American Philosophical Quarterly* 14.1 (1977), pp. 1–15. ISSN: 2152-1123.

[3] Peter Achinstein. *The Nature of Explanation*. New York, NY, USA: Oxford University Press, 1983. ISBN: 978-0-19-503743-2.

[4] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

[5] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. *Sanity Checks for Saliency Maps*. 2020. arXiv: 1810.03292.

[6] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. "Black Box Fairness Testing of Machine Learning Models". In: *Proceedings of the 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2019 (Tallinn, Estonia). Ed. by Marlon Dumas, Dietmar Pfahl, Sven Apel, and Alessandra Russo. New York, NY, USA: Association for Computing Machinery, 2019, pp. 625–635. ISBN: 978-1-4503-5572-8. DOI: 10.1145/3338906.3338937.

[7] Larry Alexander and Michael Moore. "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University, 2016. URL: https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/.

[8] Larry Alexander and Michael Moore. "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University, 2021. URL: https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.

[9] Colin Allen, Iva Smit, and Wendell Wallach. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches". In: *Ethics and Information Technology* 7 (2005), pp. 149–155. ISSN: 1572-8439. DOI: 10.1007/s10676-006-0004-4.

[10] Colin Allen, Gary Varner, and Jason Zinser. "Prolegomena to Any Future Artificial Moral Agent". In: *Journal of Experimental & Theoretical Artificial Intelligence* 12.3 (2000), pp. 251–261. ISSN: 1362-3079. DOI: 10.1080/09528130050111428.

[11] Colin Allen, Wendell Wallach, and Iva Smit. "Why Machine Ethics?" In: *IEEE Intelligent Systems* 21.4 (2006), pp. 12–17. ISSN: 1941-1294. DOI: 10.1109/MIS.2006.83.

[12] Jose M. Alonso and Gracian Trivino. "An Essay on Self-explanatory Computational Intelligence: A Linguistic Model of Data Processing Systems". In: *Proceedings of the 1st Workshop on Explainable Computational Intelligence*. XCI 2017. Ed. by Martin Pereira-Fariña and Chris Reed. Dundee, Scotland, UK: Association for Computational Linguistics, 2017. ISBN: 978-1-945626-79-1. DOI: 10.18653/v1/W17-3704.

[13] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. "Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy). Ed. by Fabio Paternò, Nuria Oliver, Cristina Conati, Lucio Davide Spano, and Nava Tintarev. IUI 2020. New York, NY, USA: Association for Computing Machinery, 2020, pp. 275–285. ISBN: 978-1-4503-7118-6. DOI: 10.1145/3377325.3377519.

[14] Maria Alvarez. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford, England, UK: Oxford University Press, 2010. ISBN: 978-0-19-955000-5. DOI: 10.1093/acprof:oso/9780199550005.001.0001.

[15] Maria Alvarez. "Reasons for Action: Justification, Motivation, Explanation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017. URL: https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.

[16] Leila Amgoud, Jean-François Bonnefon, and Henri Prade. "An Argumentation-Based Approach to Multiple Criteria Decision". In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Ed. by Lluís Godo. Vol. 3571. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2005, pp. 269–280. ISBN: 978-3-540-31888-0. DOI: 10.1007/11518655_24.

[17] Leila Amgoud and Henri Prade. "Using Arguments for Making and Explaining Decisions". In: *Artificial Intelligence* 173.3–4 (2009), pp. 413–436. ISSN: 0004-3702. DOI: 10.1016/j.artint.2008.11.006.

[18] Elvio G. Amparore, Alan Perotti, and Paolo Bajardi. "To Trust or Not to Trust an Explanation: Using LEAF to Evaluate Local Linear XAI Methods". In: *PeerJ Computer Science* 7, e479 (2021). ISSN: 2376-5992. DOI: 10.7717/peerj-cs.479.

[19] Alan Ross Anderson. "A Reduction of Deontic Logic to Alethic Modal Logic". In: *Mind* LXVII.265 (1958), pp. 100–103. ISSN: 0026-4423. DOI: 10.1093/mind/LXVII.265.100.

[20] Michael Anderson and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent". In: *AI Magazine* 28.4 (2007), pp. 15–26. ISSN: 0738-4602. DOI: 10.1609/aimag.v28i4.2065.

[21] Michael Anderson and Susan Leigh Anderson. "The Status of Machine Ethics: A Report From the AAAI Symposium". In: *Minds and Machines* 17.1 (2007), pp. 1–10. ISSN: 1572-8641. DOI: 10.1007/s11023-007-9053-7.

[22] Michael Anderson and Susan Leigh Anderson. *Machine Ethics*. Cambridge, England, UK: Cambridge University Press, 2011. ISBN: 978-0-511-97803-6. DOI: 10.1017/cbo9780511978036.

[23] Michael Anderson, Susan Leigh Anderson, and Chris Armen. "Towards Machine Ethics". In: *Proceedings of the AAAI Workshop on Agent Organizations: Theory and Practice*. AAAI WS 2004 (San Jose, California, USA). Ed. by Virginia Dignum, Daniel Corkill, Catholijn Jonker, and Frank Dignum. AAAI Technical Report WS-04-02. Palo Alto, CA, USA: AAAI Press, 2004, pp. 53–59. URL: https://aaai.org/Library/Workshops/2004/ws04-02-008.php.

[24] Michael Anderson, Susan Leigh Anderson, and Chris Armen. "An Approach to Computing Ethics". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 56–63. ISSN: 1941-1294. DOI: 10.1109/MIS.2006.64.

[25] Susan Leigh Anderson. "Asimov's "Three Laws of Robotics" and Machine Metaethics". In: *AI & Society* 22 (2008), pp. 477–493. ISSN: 1435-5655. DOI: 10.1007/s00146-007-0094-5.

[26] Susan Leigh Anderson. "How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better". In: *Machine Ethics*. Cambridge, England, UK: Cambridge University Press, 2011. Chap. 30, pp. 524–530. ISBN: 978-0-511-97803-6. DOI: 10.1017/cbo9780511978036.036.

[27] Susan Leigh Anderson. "Machine Metaethics". In: *Machine Ethics*. Cambridge, England, UK: Cambridge University Press, 2011. Chap. 2, pp. 21–27. ISBN: 978-0-511-97803-6. DOI: 10.1017/cbo9780511978036.004.

[28] Susan Leigh Anderson and Michael Anderson. "A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care". In: *Proceedings of the AAAI Workshop on Human-Robot Interaction in Elder Care*. AAAI WS 2011 (San Francisco, California, USA). Ed. by Ted Metzler. AAAI Technical Report WS-11-12. Palo Alto, CA, USA: AAAI Press, 2011, pp. 1–7. URL: https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3812/4274.

[29] Sule Anjomshoae, Kary Främling, and Amro Najjar. "Explanations of Black-Box Model Predictions by Contextual Importance and Utility". In: *Proceedings of the 1st International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. EXTRAAMAS 2019 (Montréal, Québec, Canada). Ed. by Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling. Lecture Notes in Computer Science 11763. Cham, Switzerland: Springer International Publishing, 2019, pp. 95–109. ISBN: 978-3-030-30391-4. DOI: 10.1007/978-3-030-30391-4_6.

[30] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. "Explainable Agents and Robots: Results from a Systematic Literature Review". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS 2019 (Montréal, Québec, Canada). Ed. by Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor. Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088. ISBN: 978-1-4503-6309-9. URL: http://dl.acm.org/citation.cfm?id=3331806.

[31] Gertrude E. M. Anscombe. "Modern Moral Philosophy". In: *Philosophy* 33.124 (1958), pp. 1–19. ISSN: 0031-8191. DOI: 10.1017/S0031819100037943.

[32] Gertrude E. M. Anscombe. *Intention*. Oxford, England, UK: Blackwell Press, 1962.

[33] Thomas Aquinas. *Summa Theologica*. 1485.

[34] Aristotle. *Analytica posteriora*.

[35] Aristotle. *The Nicomachean Ethics*.

[36] Cristian Arteaga. *Interpretable Machine Learning for Image Classification with LIME – Increase confidence in your machine-learning model by understanding its predictions*. 2019. URL: https://towardsdatascience.com/interpretable-machine-learning-for-image-classification-with-lime-ea947e82ca13 (visited on 10/31/2022).

[37] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. 2019. arXiv: 1909.03012.

[38] Kevin D. Ashley and Bruce M. McLaren. "Reasoning With Reasons in Case-Based Comparisons". In: *Proceedings of the 1st International Conference on Case-Based Reasoning*. ICCBR 1995 (Sesimbra, Portugal). Ed. by Manuela Veloso and Agnar Aamodt. Vol. 1010. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 1995, pp. 133–144. ISBN: 978-3-540-48446-2. DOI: 10.1007/3-540-60598-3_13.

[39] Isaac Asimov. "Runaround". In: *Astounding Science Fiction* 29.1 (1942), pp. 94–103. ISSN: 0003-2603.

[40] Isaac Asimov. *I, Robot*. New York, NY, USA: Gnome Press, 1950.

[41] Isaac Asimov. "The Duel". In: *Robots and Empire*. New York, NY, USA: Doubleday Books, 1985. ISBN: 0-385-19092-1.

[42] Nafsika Athanassoulis. "Virtue Ethics". In: *The Internet Encyclopedia of Philosophy* (2021). ISSN: 2161-0002. URL: https://www.iep.utm.edu/virtue (visited on 11/01/2021).

[43] Martin Atzmueller. "Towards Socio-Technical Design of Explicative Systems: Transparent, Interpretable and Explainable Analytics and Its Perspectives in Social Interaction Contexts information". In: *Proceedings of the 3rd Workshop on Affective Computing and Context Awareness in Ambient Intelligence*. AfCAI 2019 (Cartagena, Spain). Ed. by Grzegorz J. Nalepa, José M. Ferrández, José T. Palma-Méndez, and Vicente Julián. CEUR Workshop Proceedings 2609. CEUR-WS, 2019, pp. 1–8. URL: http://ceur-ws.org/Vol-2609/AfCAI2019_paper_9.pdf.

[44] John L. Austin. "A Plea for Excuses". In: *Ordinary Language: Essays in Philosophical Method*. Ed. by V. C. Chappell. Englewood Cliffs, NJ, USA: Prentice Hall, 1964.

[45] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Y. Ng, and Nigam H. Shah. "Improving Palliative Care With Deep Learning". In: *IEEE International Conference on Bioinformatics and Biomedicine*. BIBM 2017 (Kansas City, Missouri, USA). Ed. by Xiaohua Hu, Chi-Ren Shyu, Yana Bromberg, Jean Gao, Yang Gong, Dmitry Korkin, Illhoi Yoo, and Huiru Jane Zheng. Piscataway, NJ, USA: IEEE, 2017, pp. 311–316. ISBN: 978-1-5090-3050-7. DOI: 10.1109/BIBM.2017.8217669.

[46] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment". In: *Nature* 563.7729 (2018), pp. 59–64. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0637-6.

[47] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, M. J. Crockett, Jim A. C. Everett, Theodoros Evgeniou, Alison Gopnik, Julian C. Jamison, Tae Wan Kim, S. Matthew Liao, Michelle N. Meyer, John Mikhail, Kweku Opoku-Agyemang, Jana Schaich Borg, Juliana Schroeder, Walter Sinnott-Armstrong, Marija Slavkovik, and Josh B. Tenenbaum. "Computational Ethics". In: *Trends in Cognitive Sciences* 26.5 (2022), pp. 388–405. ISSN: 1364-6613. DOI: 10.1016/j.tics.2022.02.009.

[48] Ismaïl Baaj, Jean-Philippe Poli, and Wassila Ouerdane. "Some Insights Towards a Unified Semantic Representation of Explanation for eXplainable Artificial Intelligence". In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. NL4XAI 2019 (Tokyo, Japan). Ed. by Jose M. Alonso and Alejandro Catala. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 14–19. DOI: 10.18653/v1/W19-8404.

[49] Maria Baghramian and J. Adam Carter. "Relativism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University, 2022. URL: https://plato.stanford.edu/archives/spr2022/entries/relativism/.

[50] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. Cambridge, MA, USA: MIT Press, 2008. ISBN: 978-0-262-02649-9.

[51] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. "Transparent, Scrutable and Explainable User Models for Personalized Recommendation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 2019 (Paris, France). Ed. by Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer. New York, NY, USA: Association for Computing Machinery, 2019, pp. 265–274. ISBN: 978-1-4503-6172-9. DOI: 10.1145/3331184.3331211.

[52] Solon Barocas and Andrew D. Selbst. "Big Data's Disparate Impact". In: *California Law Review* 104 (2016), pp. 671–732. ISSN: 0008-1221. DOI: 10.2139/ssrn.2477899.

[53] James Barrat. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York, NY, USA: Thomas Dunne Books, 2013. ISBN: 978-0-312-62237-4.

[54] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Benjamins Richard, Raja Chatila, and Francisco Herrera. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012.

[55] Thomas Bartelborth. *Erklären*. Berlin, Germany: De Gruyter, 2007. ISBN: 978-3-11-091686-7. DOI: 10.1515/9783110916867.

[56] Paul Bartha. "Conditional Obligation, Deontic Paradoxes, and the Logic of Agency". In: *Annals of Mathematics and Artificial Intelligence* 9.1 (1993), pp. 1–23. ISSN: 1573-7470. DOI: 10.1007/BF01531259.

[57] Gilles Barthe, Pedro R. D'Argenio, Bernd Finkbeiner, and Holger Hermanns. "Facets of Software Doping". In: *Proceedings of the 7th International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*. ISoLA 2016 (Corfu, Greece). Ed. by Tiziana Margaria and Bernhard Steffen. Vol. 9953. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2016, pp. 601–608. ISBN: 978-3-319-47169-3. DOI: 10.1007/978-3-319-47169-3_46.

[58] William A Bauer. "Virtuous vs. Utilitarian Artificial Moral Agents". In: *AI & Society* 35.1 (2020), pp. 263–271. ISSN: 1435-5655. DOI: https://doi.org/10.1007/s00146-018-0871-3.

[59] Kevin Baum. "What the Hack is Wrong with Software Doping?" In: *Proceedings of the 7th International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*. ISoLA 2016 (Corfu, Greece). Ed. by Tiziana Margaria and Bernhard Steffen. Vol. 9953. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2016, pp. 633–647. ISBN: 978-3-319-47169-3. DOI: 10.1007/978-3-319-47169-3_49.

[60] Kevin Baum, Holger Hermanns, and Timo Speith. "From Machine Ethics to Machine Explainability and Back". In: *International Symposium on Artificial Intelligence and Mathematics*. ISAIM 2018 (Fort Lauderdale, Florida, USA). Ed. by Martin Charles, Dimitrios I. Diochnos, Jürgen Dix, Frederick Hoffman, and Guillermo R. Simari. Fort Lauderdale, FL, USA: International Symposium on Artificial Intelligence and Mathematics, 2018, pp. 1–8. URL: https://isaim2018.cs.ou.edu/papers/ISAIM2018_Ethics_Baum_etal.pdf.

[61] Kevin Baum, Holger Hermanns, and Timo Speith. "Towards a Framework Combining Machine Ethics and Machine Explainability". In: *Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology*. CREST 2018 (Thessaloniki, Greece). Ed. by Bernd Finkbeiner and Samantha Kleinberg. Sydney, NSW, AU: Electronic Proceedings in Theoretical Computer Science, 2018, pp. 34–49. DOI: 10.4204/EPTCS.286.4.

[62] Kevin Baum, Maximilian A. Köhl, and Eva Schmidt. "Two Challenges for CI Trustworthiness and How to Address Them". In: *Proceedings of the 1st Workshop on Explainable Computational Intelligence*. XCI 2017 (Santiago de Compostela, Spain). Ed. by Martin Pereira-Fariña and Chris Reed. Dundee, Scotland, UK: Association for Computational Linguistics, 2017. ISBN: 978-1-945626-79-1. DOI: 10.18653/v1/W17-3701.

[63] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. "From Responsibility to Reason-Giving Explainable Artificial Intelligence". In: *Philosophy & Technology* 35.1 (2022), pp. 1–30. ISSN: 2210-5441. DOI: 10.1007/s13347-022-00510-w.

[64] Mary Bazire and Patrick Brézillon. "Understanding Context Before Using It". In: *Proceedings of the 5th International and Interdisciplinary Conference on Modeling and Using Context*. CONTEXT 2005 (Paris, France). Ed. by Anind Dey, Boicho Kokinov, David Leake, and Roy Turner. Vol. 3554. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2005, pp. 29–40. ISBN: 978-3-540-31890-3. DOI: 10.1007/11508373_3.

[65] John Beatty. "The Evolutionary Contingency Thesis". In: *Concepts, Theories, and Rationality in the Biological Sciences*. Ed. by Gereon Wolters and James G. Lennox. 1995, pp. 45–81. ISBN: 978-0-8229-3913-9.

[66] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. *Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals*. 2018. arXiv: 1807.03418.

[67] Paul Bello and Selmer Bringsjord. "On How to Build a Moral Machine". In: *Topoi* 32.2 (2013), pp. 251–266. ISSN: 1572-8749. DOI: 10.1007/s11245-012-9129-8.

[68] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. 1789.

[69] Julien Bidot, Susanne Biundo, Tobias Heinroth, Wolfgang Minker, Florian Nothdurft, and Bernd Schattenberg. "Verbal Plan Explanations for Hybrid Planning". In: *Proceedings of the Multikonferenz Wirtschaftsinformatik*. MKWI 2010 (Göttingen, Germany). Ed. by Matthias Schumann, Lutz M. Kolbe, Michael H. Breitner, and Arne Frerichs. Göttingen, Germany: Universitätsverlag Göttingen, 2010, pp. 2309–2320. ISBN: 978-3-941875-31-9. DOI: 10.17875/gup2010-1478.

[70] Mustafa Bilgic and Raymond J. Mooney. "Explaining Recommendations: Satisfaction vs. Promotion". In: *Proceedings of the Beyond Personalization Workshop on the Next Stage of Recommender Systems Research*. Beyond Personalization@IUI 2005 (San Diego, California, USA). Ed. by Mark van Setten, Sean M. McNee, and Joseph A. Konstan. Minneapolis, MN, USA: GroupLens, 2005, pp. 13–18. URL: https://grouplens.org/beyond2005/full/bilgic.pdf.

[71] Daniel Billsus and Michael J. Pazzani. "A Personal News Agent That Talks, Learns and Explains". In: *Proceedings of the 3rd Annual Conference on Autonomous Agents*. AGENTS 1999 (Seattle, Washington, USA). Ed. by Oren Etzioni, Jörg P. Müller, and Jeffrey M. Bradshaw. New York, NY, USA: Association for Computing Machinery, 1999, pp. 268–275. ISBN: 978-1-58113-066-9. DOI: 10.1145/301136.301208.

[72] Reuben Binns, Van Kleek Max, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 36th Conference on Human Factors in Computing Systems*. CHI 2018 (Montréal, Québec, Canada). Ed. by Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox. New York, NY, USA: Association for Computing Machinery, 2018, 377, pp. 1–14. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173951.

[73] Or Biran and Courtenay Cotton. "Explanation and Justification in Machine Learning: A Survey". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 8–13. URL: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

[74] Or Biran and Kathleen R McKeown. "Human-Centric Justification of Machine Learning Predictions". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI 2017 (Melbourne, Victoria, Australia). Ed. by Carles Sierra. IJCAI Organization, 2017, pp. 1461–1467. ISBN: 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/202.

[75] Shawn A. Bohner and Robert S. Arnold. *Software Change Impact Analysis*. Washington, DC, USA: IEEE Computer Society Press, 1996. ISBN: 978-0-8186-7384-9.

[76] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. "The Social Dilemma of Autonomous Vehicles". In: *Science* 352.6293 (2016), pp. 1573–1576. ISSN: 0036-8075. DOI: 10.1126/science.aaf2654.

[77] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. "Towards Providing Explanations for AI Planner Decisions". In: *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. IJCAI/ECAI XAI 2018 (Stockholm, Sweden). Ed. by David W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni. 2018, pp. 11–17. arXiv: 1810.06338.

[78] Denny Borsboom, Angelique O. J. Cramer, and Annemarie Kalis. "Brain disorders? Not Really... Why Network Structures Block Reductionism in Psychopathology Research". In: *Behavioral and Brain Sciences* 42 (2018), pp. 1–54. ISSN: 1469-1825. DOI: 10.1017/S0140525X17002266.

[79] Nick Bostrom and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence". In: ed. by Keith Frankish and William M. Ramsey. Cambridge, England, UK: Cambridge University Press, 2014, pp. 316–334. ISBN: 978-0-521-87142-6. DOI: 10.1017/CBO9781139046855.020.

[80] David Bourget and David J. Chalmers. "What Do Philosophers Believe?" In: *Philosophical Studies* 170.3 (2014), pp. 465–500. ISSN: 1573-0883. DOI: 10.1007/s11098-013-0259-7.

[81] Gerhard Brewka, Hannes Strass, Johannes Peter Wallner, and Stefan Woltran. "Weighted Abstract Dialectical Frameworks". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI/IAAI/EAAI 2018 (New Orleans, Louisiana, USA). Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. Palo Alto, CA, USA: AAAI Press, pp. 1779–1786. ISBN: 978-1-57735-800-8. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16373.

[82] Chris Brinton. "A Framework for Explanation of Machine Learning Decisions". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W. Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 14–18.

[83] Anja Broeck, Maarten Vansteenkiste, Hans Witte, Bart Soenens, and Willy Lens. "Capturing Autonomy, Competence, and Relatedness at Work: Construction and Initial Validation of the Work-Related Basic Need Satisfaction Scale". In: *Journal of Occupational and Organizational Psychology* 83.4 (2010), pp. 981–1002. ISSN: 2044-8325. DOI: 10.1348/096317909x481382.

[84] Jan M. Broersen. "CTL.STIT: Enhancing ATL To Express Important Multi-Agent System Verification Properties". In: *AAMAS 2010* (Toronto, Ontario, Canada). Ed. by Wiebe van der Hoek, Gal A. Kaminka, Yves Lespérance, Michael Luck, and Sandip Sen. Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 683–690. ISBN: 978-0-9826571-1-9. URL: https://dl.acm.org/citation.cfm?id=1838296.

[85] Jan M. Broersen. "Deontic Epistemic Stit Logic Distinguishing Modes of Mens Rea". In: *Journal of Applied Logic* 9.2 (2011), pp. 137–152. ISSN: 1570-8683. DOI: 10.1016/j.jal.2010.06.002.

[86] Jan M. Broersen. "Probabilistic Stit Logic". In: *Proceedings of the 11th Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. ECSQARU 2011 (Belfast, Northern Ireland, UK). Ed. by Weiru Liu. Vol. 6717. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2011, pp. 521–531. ISBN: 978-3-642-22152-1. DOI: 10.1007/978-3-642-22152-1_44.

[87] Teon L. Brooks, Christopher Olah, Abhinav Prakash, and Ludwig Schubert. *Lucid Tutorial*. 2020. URL: https://github.com/tensorflow/lucid/blob/master/notebooks/tutorial.ipynb (visited on 10/31/2022).

[88] Miles Brundage. "Limitations and Risks of Machine Ethics". In: *Journal of Experimental & Theoretical Artificial Intelligence* 26.3 (2014), pp. 355–372. ISSN: 1362-3079. DOI: 10.1080/0952813X.2014.895108.

[89] Wasja Brunotte, Larissa Chazette, Verena Klös, Eric Knauss, Timo Speith, and Andreas Vogelsang. "Welcome to the First International Workshop on Requirements Engineering for Explainable Systems (RE4ES)". In: *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops*. REW 2021 (Notre Dame, Indiana, USA). Ed. by Tao Yue and Mehdi Mirakhorli. Piscataway, NJ, USA: IEEE, 2021. ISBN: 978-1-6654-1898-0. DOI: 10.1109/REW53955.2021.00028.

[90] Wasja Brunotte, Larissa Chazette, Verena Klös, and Timo Speith. "Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering". In: *Proceedings of the 28th International Working Conference on Requirements Engineering: Foundation for Software Quality*. Ed. by Vincenzo Gervasi and Andreas Vogelsang. Vol. 13216. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing, 2022, pp. 26–32. ISBN: 978-3-030-98464-9. DOI: 10.1007/978-3-030-98464-9_3.

[91] Bruce G. Buchanan and Edward H. Shortliffe. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Boston, MA, USA: Addison-Wesley, 1984. ISBN: 978-0-201-10172-0.

[92] Jenna Burrell. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms". In: *Big Data & Society* 3.1 (2016), pp. 1–12. DOI: 10.1177/2053951715622512.

[93] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems". In: *Proceedings of the 3rd International Conference on Healthcare Informatics*. ICHI 2015 (Dallas, TX, USA). Ed. by Prabhakaran Balakrishnan, Jaideep Srivatsava, Wai-Tat Fu, Sanda M. Harabagiu, and Fei Wang. Piscataway, NJ, USA: IEEE, 2015, pp. 160–169. ISBN: 978-1-4673-9548-9. DOI: 10.1109/ICHI.2015.26.

[94] Krister Bykvist. *Utilitarianism: A Guide for the Perplexed*. London, England, UK: Continuum, 2010. ISBN: 978-0-8264-9808-3.

[95] Ruth M. J. Byrne. "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. IJCAI 2019 (Macao, China). Ed. by Sarit Kraus. IJCAI Organization, 2019, pp. 6276–6282. ISBN: 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/876.

[96] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. "The Effects of Example-Based Explanations in a Machine Learning Interface". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 258–262. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302289.

[97] A. Caliskan, J. J. Bryson, and A. Narayanan. "Semantics Derived Automatically From Language Corpora Contain Human-Like Biases". In: *Science* 356 (2017), pp. 183–186. ISSN: 1095-9203. DOI: 10.1126/science.aal4230.

[98] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 2015 (Sydney, New South Wales, Australia). Ed. by Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1721–1730. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788613.

[99] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". In: *Electronics* 8.8 (2019). ISSN: 2079-9292. DOI: 10.3390/electronics8080832.

[100] Stephen Cave, Rune Nyrup, Karina Vold, and Adrian Weller. "Motivations and Risks of Machine Ethics". In: *Proceedings of the IEEE* 107.3 (2019), pp. 562–574. ISSN: 1558-2256. DOI: 10.1109/JPROC.2018.2865996.

[101] Urszula Chajewska and Joseph Y. Halpern. "Defining Explanation in Probabilistic Systems". In: *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*. UAI 1997 (Providence, Rhode Island, USA). Ed. by Dan Geiger and Prakash P. Shenoy. San Francisco, CA, USA: Morgan Kaufmann, 1997, pp. 62–71. ISBN: 978-1-55860-485-8. arXiv: 1810.06338.

[102] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. "Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior". In: *Proceedings of the 29th International Conference on Automated Planning and Scheduling*. ICAPS 2019 (Berkeley, California, USA). Ed. by J. Benton, Nir Lipovetzky, Eva Onaindia, David E. Smith, and Siddharth Srivastava. Palo Alto, CA, USA: AAAI Press, 2019, pp. 86–96. URL: https://aaai.org/ojs/index.php/ICAPS/article/view/3463.

[103] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. "Plan Explanations as Model Reconciliation". In: *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. HRI 2019 (Daegu, South Korea). Ed. by Jung Kim, Adriana Tapus, David Sirkin, Malte Jung, and Sonya S. Kwak. Piscataway, NJ, USA: IEEE, 2019, pp. 258–266. ISBN: 978-1-5386-8555-6. DOI: 10.1109/HRI.2019.8673193.

[104] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. WACV 2018 (Lake Tahoe, Nevada, USA). Piscataway, NJ, USA: IEEE, 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097.

[105] Larissa Chazette, Wasja Brunotte, and Timo Speith. "Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue". In: *Proceedings of the 29th IEEE International Requirements Engineering Conference*. RE 2021 (Notre Dame, Indiana, USA). Piscataway, NJ, USA: IEEE, 2021, pp. 197–208. ISBN: 978-1-6654-2856-9. DOI: 10.1109/RE51729.2021.00025.

[106] Larissa Chazette, Wasja Brunotte, and Timo Speith. *Supplementary Material for Research Paper "Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue"*. July 2021. DOI: 10.5281/zenodo.5114922. URL: https://doi.org/10.5281/zenodo.5114922.

[107] Larissa Chazette, Wasja Brunotte, and Timo Speith. "Explainable Software Systems: From Requirements Analysis to System Evaluation". In: *Requirements Engineering* 27.4 (2022), pp. 457–487. ISSN: 1432-010X. DOI: 10.1007/s00766-022-00393-5.

[108] Larissa Chazette, Oliver Karras, and Kurt Schneider. "Do End-Users Want Explanations? Analyzing the Role of Explainability as an Emerging Aspect of Non-Functional Requirements". In: *Proceedings of the 27th IEEE International Requirements Engineering Conference*. RE 2019 (Jeju Island, South Korea). Ed. by Daniela E. Damian, Anna Perini, and Seok-Won Lee. 2019, pp. 223–233. ISBN: 978-1-7281-3912-8. DOI: 10.1109/RE.2019.00032.

[109] Jiaoyan Chen, Freddy Lécué, Jeff Pan, Ian Horrocks, and Huajun Chen. "Knowledge-Based Transfer Learning Explanation". In: *Proceedings of the 16th International Conference for Principles of Knowledge Representation and Reasoning*. KR 2018 (Tempe, Arizona, USA). Ed. by Michael Thielscher, Francesca Toni, and Frank Wolter. Palo Alto, CA, USA: AAAI Press, 2018, pp. 349–358. ISBN: 978-1-57735-803-9. arXiv: 1807.08372.

[110] Li Chen and Feng Wang. "Explaining Recommendations Based on Feature Sentiments in Product Reviews". In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. IUI 2017 (Limassol, Cyprus). Ed. by George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu. New York, NY, USA: Association for Computing Machinery, 2017, pp. 17–28. ISBN: 978-1-4503-4348-0. DOI: 10.1145/3025171.3025173.

[111] Li Chen, Dongning Yan, and Feng Wang. "User Evaluations on Sentiment-Based Recommendation Explanations". In: *ACM Transactions on Interactive Intelligent Systems* 9.4, 20 (2019), pp. 1–38. DOI: 10.1145/3282878.

[112] Tsai-Wei Chen and S. Shyam Sundar. "This App Would Like to Use Your Current Location to Better Serve You: Importance of User Assent and System Transparency in Personalized Mobile Services". In: *Proceedings of the 36th Conference on Human Factors in Computing Systems*. CHI 2018 (Montréal, Québec, Canada). Ed. by Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox. New York, NY, USA: Association for Computing Machinery, 2018, 537, pp. 1–13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174111.

[113] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. "Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders". In: *Proceedings of the 37th Conference on Human Factors in Computing Systems*. CHI 2019 (Glasgow, Scotland, UK). Ed. by Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos. New York, NY, USA: Association for Computing Machinery, 2019, 559, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300789.

[114] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems". In: *Joint Proceedings of the 24th ACM Conference on Intelligent User Interfaces Workshops*. IUI WS 2019 (Los Angeles, California, USA). Ed. by Christoph Trattner, Denis Parra, and Nathalie Riche. CEUR Workshop Proceedings 2327. CEUR-WS, 2019, pp. 1–6. URL: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-7.pdf.

[115] Romane Clark and Paul Welsh. *Introduction to Logic*. New York, NY, USA: Van Nostrand Company, 1962. ISBN: 978-0-4420-1571-8.

[116] Miruna-Adriana Clinciu and Helen Hastie. "A Survey of Explainable AI Terminology". In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. NL4XAI 2019 (Tokyo, Japan). Ed. by Jose M. Alonso and Alejandro Catala. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 8–13. DOI: 10.18653/v1/W19-8403.

[117] Christopher Cloos. "The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism". In: *Proceedings of the AAAI Fall Symposium on Machine Ethics*. AAAI FS 2005 (Arlington, Virginia, USA). Ed. by Michael Anderson, Susan Leigh Anderson, and Chris Armen. AAAI Technical Report FS-05-06. Palo Alto, CA, USA: AAAI Press, 2005, pp. 38–45. ISBN: 978-1-57735-252-5. URL: https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-006.pdf.

[118] Jérémie Clos, Nirmalie Wiratunga, and Stewart Massie. "Towards Explainable Text Classification by Jointly Learning Lexicon and Modifier Terms". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 19–23.

[119] Mark Coeckelbergh. "The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics". In: *Philosophy & Technology* 27.1 (2014), pp. 61–77. ISSN: 2210-5441. DOI: 10.1007/s13347-013-0133-8.

[120] Mark Coeckelbergh and David J. Gunkel. "Facing Animals: A Relational, Other-Oriented Approach to Moral Standing". In: *Journal of Agricultural and Environmental Ethics* 27.5 (2014), pp. 715–733. ISSN: 1187-7863. DOI: 10.1007/s10806-013-9486-3.

[121] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. "Moral Decision Making Frameworks for Artificial Intelligence". In: *International Symposium on Artificial Intelligence and Mathematics*. ISAIM 2018 (Fort Lauderdale, Florida, USA). Ed. by Martin Charles, Dimitrios I. Diochnos, Jürgen Dix, Frederick Hoffman, and Guillermo R. Simari. Fort Lauderdale, FL, USA: International Symposium on Artificial Intelligence and Mathematics, 2018. URL: http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Ethics_Conitzer_etal.pdf.

[122] Stephen A. Cook. "The Complexity of Theorem-Proving Procedures". In: *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*. STOC 1971 (Shaker Heights, Ohio, USA). Ed. by Michael A. Harrison, Ranan B. Banerji, and Jeffrey D. Ullman. New York, NY, USA: Association for Computing Machinery, 1971, pp. 151–158. ISBN: 978-1-4503-7464-4. DOI: 10.1145/800157.805047.

[123] Kelley Cotter, Janghee Cho, and Emilee J. Rader. "Explaining the News Feed Algorithm: An Analysis of the 'News Feed FYI' Blog". In: *Proceedings of the 35th Conference on Human Factors in Computing Systems Extended Abstracts*. CHI EA 2017. Ed. by Gloria Mark, Susan R. Fussell, Cliff Lampe, M. C. Schraefel, Juan P. Hourcade, Caroline Appert, and Daniel Wigdor. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1553–1560. ISBN: 978-1-4503-4656-6. DOI: 10.1145/3027063.3053114.

[124] Henriette S. M. Cramer, Vanessa Evers, Satyan Ramlal, van Someren Maarten, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob J. Wielinga. "The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender". In: *User Modeling and User-Adapted Interaction* 18.5 (2008), pp. 455–496. ISSN: 1573-1391. DOI: 10.1007/s11257-008-9051-3.

[125] Carl F. Craver. *Explaining the Brain*. Oxford, England, UK: Oxford University Press, 2007. ISBN: 978-0-19-929931-7. DOI: 10.1093/acprof:oso/9780199299317.001.0001.

[126] Barnaby Crook, Maximilian Schlüter, and Timo Speith. "Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)". In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*. REW 2023 (Hannover, Germany). Ed. by Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider. Piscataway, NJ, USA: IEEE, 2023.

[127] Luiz M. Cysneiros, Majid Raffi, and Julio C. S. do Prado Leite. "Software Transparency as a Key Requirement for Self-Driving Cars". In: *Proceedings of the 26th IEEE International Requirements Engineering Conference*. RE 2018 (Banff, Alberta, Canada). Ed. by Guenther Ruhe, Walid Maalej, and Daniel Amyot. Piscataway, NJ, USA: IEEE, 2018, pp. 382–387. ISBN: 978-1-5386-7418-5. DOI: 10.1109/RE.2018.00-21.

[128] Luiz M. Cysneiros and Vera M. B. Werneck. "An Initial Analysis on How Software Transparency and Trust Influence each other". In: *Anais do Workshop em Engenharia de Requisitos*. WER 2009 (Valparaíso, Chile). Ed. by Claudia P. Ayala, Carla T. L. L. Silva, and Hernán Astudillo. 2009. URL: http://wer.inf.puc-rio.br/WERpapers/artigos/artigos_WER09/cysneiros.pdf.

[129] Pedro R. D'Argenio, Gilles Barthe, Sebastian Biewer, Bernd Finkbeiner, and Holger Hermanns. "Is Your Software on Dope? – Formal Analysis of Surreptitiously "Enhanced" Programs". In: *Proceedings of the 26th European Symposium on Programming*. ESOP 2017 (Uppsala, Sweden). Vol. 10201. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2017, pp. 83–110. ISBN: 978-3-662-54434-1. DOI: 10.1007/978-3-662-54434-1_4.

[130] Erik S. Dahl. "Appraising Black-Boxed Technology: The Positive Prospects". In: *Philosophy & Technology* 31.4 (2018), pp. 571–591. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0275-1.

[131] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. "Explainable Software Analytics". In: *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*. ICSE NIER 2018 (Gothenburg, Sweden). Ed. by Andrea Zisman and Sven Apel. New York, NY, USA: Association for Computing Machinery, 2018, pp. 53–56. ISBN: 978-1-4503-5662-6. DOI: 10.1145/3183399.3183424.

[132] John Danaher. "The Rise of the Robots and the Crisis of Moral Patiency". In: *AI & Society* 34.1 (2019), pp. 129–136. ISSN: 1435-5655. DOI: 10.1007/s00146-017-0773-9.

[133] Jonathan Dancy. *Practical Reality*. New York, NY, USA: Oxford University Press, 2000. ISBN: 978-0-19-151961-1. DOI: 10.1093/0199253056.001.0001.

[134] Jonathan Dancy. *Ethics Without Principles*. New York, NY, USA: Oxford University Press, 2004. ISBN: 978-0-19-929768-9. DOI: 10.1093/0199270023.001.0001.

[135] Jonathan Dancy. "Moral Particularism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017. URL: https://plato.stanford.edu/archives/win2017/entries/moral-particularism/.

[136] Kate Darling. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects". In: *Robot Law*. Ed. by Ryan Calo, A. Michael Froomkin, and Ian Kerr. Cheltenham, England, UK: Edward Elgar Publishing, 2016. Chap. 9, pp. 213–232. ISBN: 978-1-78347-672-5. DOI: 10.4337/9781783476732.00017.

[137] Keith Darlington. "Aspects of Intelligent Systems Explanation". In: *Universal Journal of Control and Automation* 1.2 (2013), pp. 40–51. ISSN: 2331-6500. DOI: 10.13189/ujca.2013.010204.

[138] Jeffrey Dastin. *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on 01/30/2022).

[139] Anupam Datta, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments With Learning Systems". In: *Proceedings of the 37th IEEE Symposium on Security and Privacy*. SP 2016 (San Jose, California, USA). Ed. by Michael Locasto, Vitaly Shmatikov, and Erlingsson Úlfar. Piscataway, NJ, USA: IEEE, 2016, pp. 598–617. ISBN: 978-1-5090-0824-7. DOI: 10.1109/SP.2016.42.

[140] David Davenport. "Moral Mechanisms". In: *Philosophy & Technology* 27.1 (2014), pp. 47–60. ISSN: 2210-5441. DOI: 10.1007/s13347-013-0147-2.

[141] Donald Davidson. "Actions, Reasons, and Causes". In: *The Journal of Philosophy* 60.23 (1963), pp. 685–700. ISSN: 1939-8549. DOI: 10.2307/2023177.

[142] Morteza Dehghani, Emmett Tomai, Kenneth D Forbus, and Matthew Klenk. "An Integrated Reasoning Approach to Moral Decision-Making". In: *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*. AAAI 2008 (Chicago, Illinois, USA). Ed. by Dieter Fox and Carla P. Gomes. Palo Alto, CA, USA: AAAI Press, 2008, pp. 1280–1286. ISBN: 978-1-57735-368-3. URL: https://www.aaai.org/Library/AAAI/2008/aaai08-203.php.

[143] Kevin M. DeLapp. "Metaethics". In: *The Internet Encyclopedia of Philosophy* (2021). ISSN: 2161-0002. URL: https://www.iep.utm.edu/metaethi (visited on 11/01/2021).

[144] Daniel Clement Dennett. "Three Kinds of Intentional Psychology". In: *The Intentional Stance*. Cambridge, MA, USA: MIT Press, 1987, pp. 43–68. ISBN: 978-0-262-54053-7.

[145] Louise Dennis and Michael Fisher. "Practical Challenges in Explicit Ethical Machine Reasoning". In: *International Symposium on Artificial Intelligence and Mathematics*. ISAIM 2018 (Fort Lauderdale, Florida, USA). Ed. by Martin Charles, Dimitrios I. Diochnos, Jürgen Dix, Frederick Hoffman, and Guillermo R. Simari. Fort Lauderdale, FL, USA: International Symposium on Artificial Intelligence and Mathematics, 2018. URL: http://isaim2018.cs.virginia.edu/papers/ISAIM2018%5C_Ethics%5C_Dennis%5C_Fischer.pdf.

[146] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. "Formal Verification of Ethical Choices in Autonomous Systems". In: *Robotics and Autonomous Systems* 77 (2016), pp. 1–14. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2015.11.012.

[147] Irene-Anna N. Diakidoy, Loizos Michael, and Antonis Kakas. "Knowledge Activation in Story Comprehension". In: *Journal of Cognitive Science* 18.4 (2017), pp. 439–471. ISSN: 1976-6939.

[148] Franz Dietrich and Christian List. "What Matters and How It Matters: A Choice-Theoretic Representation of Moral Theories". In: *Philosophical Review* 126.4 (2017), pp. 421–479. ISSN: 0031-8108. DOI: 10.1215/00318108-4173412.

[149] Joel Dittmer. "Applied Ethics". In: *The Internet Encyclopedia of Philosophy* (2021). ISSN: 2161-0002. URL: https://www.iep.utm.edu/ap-ethics (visited on 11/01/2021).

[150] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. "Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 275–285. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302310.

[151] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. "The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 408–416. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302274.

[152] Vicente Dominguez, Pablo Messina, Christoph Trattner, and Denis Parra. "Towards Explanations for Visual Recommender Systems of Artistic Images". In: *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. IntRS@RecSys 2018 (Vancouver, British Columbia, Canada). Ed. by Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, John O'Donovan, Giovanni Semeraro, and Martijn C. Willemsen. CEUR Workshop Proceedings 2225. CEUR-WS, 2018, pp. 69–73. URL: http://ceur-ws.org/Vol-2225/paper10.pdf.

[153] Derek Doran, Sarah Schulz, and Tarek R. Besold. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives". In: *Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML*. CEx 2017 (Bari, Italy). Ed. by Tarek R. Besold and Oliver Kutz. Vol. 2071. CEUR Workshop Proceedings. CEUR-WS, 2017, pp. 1–8. URL: http://ceur-ws.org/Vol-2071/CExAIIA%5C_2017%5C_paper%5C_2.pdf.

[154] John M. Doris. "Persons, Situations, and Virtue Ethics". In: *Noûs* 32.4 (1998), pp. 504–530. ISSN: 1468-0068. DOI: 10.1111/0029-4624.00136.

[155] Finale Doshi-Velez and Been Kim. *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608.

[156] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. *Accountability of AI Under the Law: The Role of Explanation*. 2017. arXiv: 1711.01134.

[157] Paul Dourish. "What We Talk About When We Talk About Context". In: *Personal and Ubiquitous Computing* 8.1 (2004), pp. 19–30. ISSN: 1617-4917. DOI: 10.1007/s00779-003-0253-8.

[158] Phil Dowe. *Physical Causation*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge, England, UK: Cambridge University Press, 2000. ISBN: 978-0-521-03975-8. DOI: 10.1017/CBO9780511570650.

[159] Julia Driver. "Moral Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022. URL: https://plato.stanford.edu/archives/fall2022/entries/moral-theory/.

[160] Jakob Droste, Verena Klös, Mersedeh Sadeghi, Maike Schwammberger, and Timo Speith. "Welcome to the Thirs International Workshop on Requirements Engineering for Explainable Systems (RE4ES)". In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*. REW 2023 (Hannover, Germany). Ed. by Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider. Piscataway, NJ, USA: IEEE, 2023.

[161] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. "Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI 2020 (Cagliari, Italy). Ed. by Fabio Paternò, Nuria Oliver, Cristina Conati, Lucio D. Spano, and Nava Tintarev. New York, NY, USA: Association for Computing Machinery, 2020, pp. 297–307. ISBN: 978-1-4503-7118-6. DOI: 10.1145/3377325.3377501.

[162] Ding-Zhu Du and Ker-I Ko. *Theory of Computational Complexity*. Hobocken, NJ, USA: John Wiley & Sons, 2000. ISBN: 978-0-471-34506-0. DOI: 10.1002/9781118032916.

[163] Phan Minh Dung. "On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games". In: *Artificial Intelligence* 77.2 (1995), pp. 321–357. ISSN: 0004-3702. DOI: 10.1016/0004-3702(94)00041-X.

[164] Lilian Edwards and Michael Veale. "Slave to the Algorithm: Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For". In: *Duke Law & Technology Review* 16 (2017), pp. 18–84. URL: https://scholarship.law.duke.edu/dltr/vol16/iss1/2.

[165] Kate Ehrlich, Susanna E. Kirk, John Patterson, Jamie C. Rasmussen, Steven I. Ross, and Daniel M. Gruen. "Taking Advice from Intelligent Systems: The Double-Edged Sword of Explanations". In: *Proceedings of the 16th International Conference on Intelligent User Interfaces*. IUI 2011 (Palo Alto, California, USA). Ed. by Pearl Pu, Michael J. Pazzani, Elisabeth André, and Doug Riecken. New York, NY, USA: Association for Computing Machinery, 2011, pp. 125–134. ISBN: 978-1-4503-0419-1. DOI: 10.1145/1943403.1943424.

[166] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES 2018. Ed. by Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi. New York, NY, USA: Association for Computing Machinery, 2018, pp. 81–87. ISBN: 978-1-4503-6012-8. DOI: 10.1145/3278721.3278736.

[167] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. "Automated rationale generation: a technique for explainable AI and its effects on human perceptions". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 263–274. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3278721.3278736.

[168] Thomas Eiter, Zeynep G. Saribatur, and Peter Schüller. "Abstraction for Zooming-In to Unsolvability Reasons of Grid-Cell Problems". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 7–13. arXiv: 1909.04998.

[169] E. Allen Emerson and Joseph Y. Halpern. ""Sometimes" and "Not Never" Revisited: On Branching Versus Linear Time Temporal Logic". In: *Journal of the ACM* 33.1 (1986), pp. 151–178. ISSN: 0004-5411. DOI: 10.1145/4904.4999.

[170] Mica R. Endsley. "From Here to Autonomy". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59.1 (2017), pp. 5–27. ISSN: 1547-8181. DOI: 10.1177/0018720816681350.

[171] Adrian Erasmus, Tyler D. P. Brunet, and Eyal Fisher. "What Is Interpretability?" In: *Philosophy & Technology* 34.4 (2021), pp. 833–862. ISSN: 2210-5441. DOI: 10.1007/s13347-020-00435-2.

[172] EU High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[173] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. "Transparency You Can Trust: Transparency Requirements for Artificial Intelligence Between Legal Norms and Contextual Concerns". In: *Big Data & Society* 6.1 (2019), pp. 1–14. ISSN: 2053-9517. DOI: 10.1177/2053951719860542.

[174] James Fieser. "Ethics". In: *The Internet Encyclopedia of Philosophy* (2021). ISSN: 2161-0002. URL: https://www.iep.utm.edu/ethics (visited on 11/01/2021).

[175] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". In: *Minds and Machines* 28.4 (2018), pp. 689–707. ISSN: 1572-8641. DOI: 10.1007/s11023-018-9482-5.

[176] Philippa Foot. "The Problem of Abortion and the Doctrine of Double Effect". In: *The Oxford Review* 5 (1967), pp. 5–15.

[177] Maria Fox, Derek Long, and Daniele Magazzeni. "Explainable Planning". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 24–30. arXiv: 1709.10256.

[178] Benjamin Franklin. "Letter to Dr. J. B. Priestley, 1772". In: *The Works of Benjamin Franklin*. Ed. by John Bigelow. Vol. 5. New York, NY, USA: Putnam, 1904. URL: http://oll.libertyfund.org/titles/2459#lf1438-05_head_148.

[179] Alex A Freitas. "Comprehensible Classification Models: A Position Paper". In: *SIGKDD Explorations Newsletter* 15.1 (2014), pp. 1–10. ISSN: 1931-0145. DOI: 10.1145/2594473.2594475.

[180] Gerhard Friedrich and Markus Zanker. "A Taxonomy for Generating Explanations in Recommender Systems". In: *AI Magazine* 32.3 (2011), pp. 90–98. ISSN: 0738-4602. DOI: 10.1609/aimag.v32i3.2365.

[181] Nico H. Frijda. *The Emotions*. Cambridge, England, UK: Cambridge University Press, 1986. ISBN: 978-0-521-31600-2.

[182] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. "Autonomous Self-Explanation of Behavior for Interactive Reinforcement Learning Agents". In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI 2017 (Bielefeld, Germany). Ed. by Britta Wrede, Yukie Nagai, Takanori Komatsu, Marc Hanheide, and Lorenzo Natale. New York, NY, USA: Association for Computing Machinery, 2017, pp. 97–101. ISBN: 978-1-4503-5113-3. DOI: 10.1145/3125739.3125746.

[183] Richard Gall. *Machine Learning Explainability vs. Interpretability: Two Concepts That Could Help Restore Trust in AI*. 2018. URL: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html (visited on 01/30/2022).

[184] Patrick Gamez, Daniel B. Shank, Carson Arnold, and Mallory North. "Artificial Virtue: The Machine Question and Perceptions of Moral Character in Artificial Moral Agents". In: *AI & Society* 35.4 (2020), pp. 795–809. ISSN: 1435-5655. DOI: https://doi.org/10.1007/s00146-020-00977-1.

[185] Stephen M. Gardiner. "A Core Precautionary Principle". In: *Journal of Political Philosophy* 14.1 (2006), pp. 33–60. ISSN: 0963-8016. DOI: 10.1111/j.1467-9760.2006.00237.x.

[186] James Garson. "Modal Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021. URL: https://plato.stanford.edu/archives/fall2018/entries/logic-modal/.

[187] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. "Interpretable Classification Rules in Relaxed Logical Form". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 14–20.

[188] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining Explanations: An Overview of Interpretability of Machine Learning". In: *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics*. DSAA 2018 (Turin, Italy). Ed. by Francesco Bonchi, Foster J. Provost, Tina Eliassi-Rad, Wei Wang, Ciro Cattuto, and Rayid Ghani. Piscataway, NJ, USA: IEEE, 2018, pp. 80–89. ISBN: 978-1-5386-5090-5. DOI: 10.1109/DSAA.2018.00018.

[189] Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. "Explaining Explanations to Society". In: *NIPS Workshop on Ethical, Social and Governance Issues in AI* (Montréal, Québec, Canada). 2018, pp. 1–6. arXiv: 1901.06560.

[190] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. "Toward Establishing Trust in Adaptive Agents". In: *Proceedings of the 13th International Conference on Intelligent User Interfaces*. IUI 2008 (Maspalomas, Gran Canaria, Spain). Ed. by Jeffrey M. Bradshaw, Henry Lieberman, and Steffen Staab. New York, NY, USA: Association for Computing Machinery, 2008, pp. 227–236. ISBN: 978-1-59593-987-6. DOI: 10.1145/1378773.1378804.

[191] Antoni Gomila and Alberto Amengual. "Moral Emotions for Autonomous Agents". In: *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. Ed. by Jordi Vallverdú and David Casacuberta. Hershey, PA, USA: IGI Global, 2009. Chap. 10, pp. 166–179. ISBN: 978-1-60566-354-8. DOI: 10.4018/978-1-60566-354-8.ch010.

[192] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *Proceedings of the 3rd International Conference on Learning Representations*. ICLR 2015 (San Diego, California, USA). Ed. by Yoshua Bengio and Yann LeCun. 2015. arXiv: 1412.6572.

[193] Bryce Goodman and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'". In: *AI Magazine* 38.3 (2017), pp. 50–57. ISSN: 0738-4602. DOI: 10.1609/aimag.v38i3.2741.

[194] Google PAIR. *PAIR SALIENCY – Framework-agnostic implementation for state-of-the-art saliency methods (XRAI, BlurIG, SmoothGrad, and more)*. The code is available in a Jupyter Notebook. 2022. URL: https://pair-code.github.io/saliency/ (visited on 10/31/2022).

[195] Chris Gowans. "Moral Relativism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University, 2021. URL: https://plato.stanford.edu/archives/spr2021/entries/moral-relativism/.

[196] Maartje M. A. de Graaf and Bertram F. Malle. “How People Explain Action (and Autonomous Intelligent Systems Should Too)”. In: *Proceedings of the 2017 AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction* (Arlington, Virginia, USA). Palo Alto, CA, USA: AAAI Press, 2017, pp. 19–26. URL: https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009.

[197] Christopher Grau. “There Is No "I" in "Robot": Robots and Utilitarianism”. In: *IEEE Intelligent Systems* 21.4 (2006), pp. 52–55. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.81.

[198] Stephen J. Green, Paul Lamere, Jeffrey Alexander, François Maillet, Susanna Kirk, Jessica Holt, Jackie Bourque, and Xiao-Wen Mak. “Generating Transparent, Steerable Recommendations from Textual Descriptions of Items”. In: *Proceedings of the 3rd ACM Conference on Recommender Systems*. RecSys 2009 (New York City, New York, USA). Ed. by Lawrence D. Bergman, Alexander Tuzhilin, Robin D. Burke, Alexander Felfernig, and Lars Schmidt-Thieme. New York, NY, USA: Association for Computing Machinery, 2009, pp. 281–284. ISBN: 978-1-60558-435-5. DOI: 10.1145/1639714.1639768.

[199] Shirley Gregor and Izak Benbasat. “Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice”. In: *MIS Quarterly* 23.4 (1999), pp. 497–530. ISSN: 0276-7783. DOI: 10.2307/249487.

[200] Marcello Guarim. “Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism-Generalism Debate”. In: *Machine Ethics*. Ed. by Michael Anderson and Susan Leigh Anderson. New York, NY, USA: Cambridge University Press, 2011, pp. 316–334. ISBN: 978-1-108-46175-7.

[201] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5, 93 (2019), pp. 1–42. ISSN: 0360-0300. DOI: 10.1145/3236009.

[202] David J. Gunkel. “A Vindication of the Rights of Machines”. In: *Philosophy & Technology* 27.1 (2014), pp. 113–133. ISSN: 2210-5441. DOI: 10.1007/s13347-013-0121-z.

[203] Francisco Gutiérrez, Sven Charleer, Robin de Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. “Explaining and Exploring Job Recommendations: A User-Driven Approach for Interacting with Knowledge-Based Job Recommender Systems”. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. RecSys 2019 (Copenhagen, Denmark). Ed. by Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk. New York, NY, USA: Association for Computing Machinery, 2019, pp. 60–68. ISBN: 978-1-4503-6243-6. DOI: 10.1145/3298689.3347001.

[204] William Haines. “Consequentialism”. In: *The Internet Encyclopedia of Philosophy* (). ISSN: 2161-0002. URL: https://www.iep.utm.edu/conseque (visited on 11/01/2021).

[205] John S. Hall. “Ethics for Self-Improving Machines”. In: *Machine Ethics*. Ed. by Michael Anderson and Susan Leigh Anderson. New York, NY, USA: Cambridge University Press, 2011, pp. 512–523. ISBN: 978-1-108-46175-7.

[206] Mark Hall, Daniel Harborne, Richard Tomsett, Vedran Galetic, Santiago Quintana-Amate, Alistair Nottle, and Alun Preece. “A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems”. In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 21–27.

[207] Joseph Y. Halpern and Judea Pearl. “Causes and Explanations: A Structural-Model Approach”. Part I: Causes. In: *The British Journal for the Philosophy of Science* 56.4 (2005), pp. 843–887. ISSN: 0007-0882. DOI: 10.1093/bjps/axi147.

[208] Joseph Y. Halpern and Judea Pearl. “Causes and Explanations: A Structural-Model Approach”. Part II: Explanations. In: *The British Journal for the Philosophy of Science* 56.4 (2005), pp. 889–911. ISSN: 0007-0882. DOI: 10.1093/bjps/axi148.

[209] Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. “A Study into Preferred Explanations of Virtual Agent Behavior”. In: *Proceedings of the 9th International Conference on Intelligent Virtual Agents*. IVA 2009 (Amsterdam, The Netherlands). Ed. by Zsófia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Högni Vilhjálmsson. Lecture Notes in Computer Science 5773. Berlin/Heidelberg, Germany: Springer, 2009, pp. 132–145. ISBN: 978-3-642-04380-2. DOI: 10.1007/978-3-642-04380-2_17.

[210] Robert Purves Hardie and Russell Kerr Gaye. "Physics by Aristotle". In: *The Works of Aristotle*. Ed. by Robert Maynard Hutchins. Chicago, IL, USA: William Benton, 1952, pp. 257–355. URL: http://classics.mit.edu/Aristotle/physics.2.ii.html.

[211] Richard Mervyn Hare. *Moral Thinking: Its Levels, Method, and Point*. New York, NY, USA: Oxford University Press, 1981. ISBN: 978-0-198-24660-2. DOI: 10.1093/0198246609.001.0001.

[212] Tim Harford. *Messy: How to Be Creative and Resilient in a Tidy-Minded World*. Boston, MA, USA: Little, Brown Book Group, 2016. ISBN: 978-0-349-14114-5.

[213] Gilbert Harman. "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error". In: *Proceedings of the Aristotelian Society* 99.1 (1999), pp. 315–331. ISSN: 0066-7374. DOI: 10.1111/1467-9264.00062.

[214] Renate Häuslschmid, Max von Buelow, Bastian Pfleging, and Andreas Butz. "Supporting Trust in Autonomous Driving". In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. IUI 2017 (Limassol, Cyprus). Ed. by George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu. New York, NY, USA: Association for Computing Machinery, 2017, pp. 319–329. ISBN: 978-1-4503-4348-0. DOI: 10.1145/3025171.3025198.

[215] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. "TriRank: Review-Aware Explainable Recommendation by Modeling Aspects". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM 2015 (Melbourne, Victoria, Australia). Ed. by James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1661–1670. ISBN: 978-1-4503-3794-6. DOI: 10.1145/2806416.2806504.

[216] Matthias Hein and Maksym Andriushchenko. "Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS 2017 (Long Beach, California, USA). Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. New York, NY, USA: Curran Associates, Inc., 2017, pp. 2266–2276. URL: https://proceedings.neurips.cc/paper/2017/hash/e077e1a544eec4f0307cf5c3c721d944-Abstract.html.

[217] Carl G. Hempel. "Deductive-Nomological Explanation". In: *Aspects of Scientific Explanation*. New York, NY, USA: Free Press, 1965, pp. 335–376. ISBN: 978-0-02-914340-7.

[218] Carl G. Hempel and Paul Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of Science* 15.2 (1948), pp. 135–175. ISSN: 1539-767X. DOI: 10.1086/286983.

[219] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. "Generating Visual Explanations". In: *Proceedings of the 14th European Conference on Computer Vision*. ECCV 2016 (Amsterdam, The Netherlands). Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Lecture Notes in Computer Science 9908. Part IV. Cham, Switzerland: Springer International Publishing, 2016, pp. 3–19. ISBN: 978-3-319-46493-0. DOI: 10.1007/978-3-319-46493-0_1.

[220] Monika Hengstler, Ellen Enkel, and Selina Duelli. "Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices". In: *Technological Forecasting and Social Change* 105 (2016), pp. 105–120. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2015.12.014.

[221] Clément Henin and Le Métayer Daniel. "Towards a Generic Framework for Black-Box Explanation Methods". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 28–34. URL: https://hal.inria.fr/hal-03127923/document.

[222] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. "Explaining Collaborative Filtering Recommendations". In: *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. CSCW 2000 (Philadelphia, Pennsylvania, USA). Ed. by Wendy A. Kellogg and Steve Whittaker. New York, NY, USA: Association for Computing Machinery, 2000, pp. 241–250. ISBN: 978-1-58113-222-9. DOI: 10.1145/358916.358995.

[223] Pamela Hieronymi. "XIV–Reasons for Action". In: *Proceedings of the Aristotelian Society* 111.3 (2011), pp. 407–427. ISSN: 0066-7374. DOI: 10.1111/j.1467-9264.2011.00316.x.

[224] Kenneth Einar Himma. "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?" In: *Ethics and Information Technology* 11.1 (2009), pp. 19–29. ISSN: 1572-8439. DOI: 10.1007/s10676-008-9167-5.

[225] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan N. Ramamurthy, and Kush R. Varshney. "TED: Teaching AI to Explain Its Decisions". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES 2019 (Honolulu, Hawaii, USA). Ed. by Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor. New York, NY, USA: Association for Computing Machinery, 2019, pp. 123–129. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314273.

[226] Christopher Read Hitchcock. "The Role of Contrast in Causal and Explanatory Claims". In: *Synthese* 107.3 (1996), pp. 395–419. ISSN: 1573-0964. DOI: 10.1007/BF00413843.

[227] Justin Chun-Ting Ho. "How Biased Is the Sample? Reverse Engineering the Ranking Algorithm of Facebook's Graph Application Programming Interface". In: *Big Data & Society* 7.1 (2020), pp. 1–15. ISSN: 2053-9517. DOI: 10.1177/2053951720905874.

[228] David H. Hodgson. *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory*. Oxford, England, UK: Clarendon Press, 1967.

[229] Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, and Maarten de Rijke. "Do News Consumers Want Explanations for Personalized News Rankings?" In: *Proceedings of the 1st FATREC Workshop on Responsible Recommendation* (Como, Italy). Ed. by Michael Ekstrand and Amit Sharma. Boise, ID, USA: Boise State ScholarWorks, 2017, pp. 1–6. DOI: 10.18122/B24D7N.

[230] Kevin Anthony Hoff and Masooda Bashir. "Trust in Automation". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57.3 (2014), pp. 407–434. ISSN: 1547-8181. DOI: 10.1177/0018720814547570.

[231] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. "Explaining Explanation For 'Explainable AI'". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62.1 (2018), pp. 197–201. ISSN: 1071-1813. DOI: 10.1177/1541931218621047.

[232] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. *Metrics for Explainable AI: Challenges and Prospects*. arXiv: 1812.04608.

[233] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. "Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models". In: *Proceedings of the 37th Conference on Human Factors in Computing Systems*. CHI 2019 (Glasgow, Scotland, UK). Ed. by Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos. New York, NY, USA: Association for Computing Machinery, 2019, 579, pp. 1–13. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300809.

[234] Joana Hois, Dimitra Theofanou-Fuelbier, and Alischa Janine Junk. "How to Achieve Explainability and Transparency in Human AI Interaction". In: *Proceedings of the 21st International Conference on Human-Computer Interaction – Posters*. HCI 2019 (Orlando, Florida, USA). Ed. by Constantine Stephanidis. Communications in Computer and Information Science 1033. PartII. Cham, Switzerland: Springer International Publishing, 2019, pp. 177–183. ISBN: 978-3-030-23528-4. DOI: 10.1007/978-3-030-23528-4_25.

[235] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. "User Trust in Intelligent Systems: A Journey Over Time". In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. IUI 2016 (Sonoma, California, USA). Ed. by Jeffrey Nichols, Jalal Mahmud, John O'Donovan, Cristina Conati, and Massimo Zancanaro. New York, NY, USA: Association for Computing Machinery, 2016, pp. 164–168. ISBN: 978-1-4503-4137-0. DOI: 10.1145/2856767.2856811.

[236] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. "Causability and explainability of artificial intelligence in medicine". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4, e1312 (2019), pp. 1–13. ISSN: 1942-4795. DOI: 10.1002/widm.1312.

[237] Helmut Horacek. "Requirements for Conceptual Representations of Explanations and How Reasoning Systems Can Serve Them". In: *Proceedings of the 1st Workshop on Explainable Computational Intelligence*. XCI 2017 (Santiago de Compostela, Spain). Ed. by Martin Pereira-Fariña and Chris Reed. Dundee, Scotland, UK: Association for Computational Linguistics, 2017. ISBN: 978-1-945626-79-1. DOI: 10.18653/v1/W17-3703.

[238] John F. Horty. *Agency and Deontic Logic*. Oxford, England, UK: Oxford University Press, 2001. ISBN: 978-0-19-513461-2. DOI: 10.1093/0195134613.001.0001.

[239] John F. Horty. *Reasons as Defaults*. Oxford, England, UK: Oxford University Press, 2012. ISBN: 978-0-19-974407-7. DOI: 10.1093/acprof:oso/9780199744077.001.0001.

[240] John F. Horty and Eric Pacuit. "Action Types in Stit Semantics". In: *The Review of Symbolic Logic* 10.4 (2017), pp. 617–637. ISSN: 1755-0203. DOI: 10.1017/S1755020317000016.

[241] Mahmood Hosseini, Alimohammad Shahri, Keith Phalp, and Raian Ali. "Four Reference Models for Transparency Requirements in Information Systems". In: *Requirements Engineering* 23.2 (2018), pp. 251–275. ISSN: 1432-010X. DOI: 10.1007/s00766-017-0265-y.

[242] Jeroen van den Hoven and Gert-Jan Lokhorst. "Deontic Logic and Computer-Supported Computer Ethics". In: *Metaphilosophy* 33.3 (2002), pp. 376–386. ISSN: 1467-9973. DOI: https://doi.org/10.1111/1467-9973.00233.

[243] Don A. Howard and Ioan Muntean. "A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA)". In: *Proceedings of the AAAI Spring Symposium on Ethical and Moral Considerations in Non-Human Agents*. AAAI SS 2016 (Palo Alto, California, USA). Ed. by Bipin Indurkhya and Georgi Stojanov. AAAI Technical Report SS-16-04. Palo Alto, CA, USA: AAAI Press, 2016. ISBN: 978-1-57735-754-4. URL: http://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12760.

[244] Don A. Howard and Ioan Muntean. "Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency". In: *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*. Ed. by Thomas M. Powers. Philosophical Studies Series 128. Cham, Switzerland: Springer International Publishing, 2017, pp. 121–159. ISBN: 978-3-319-61043-6. DOI: 10.1007/978-3-319-61043-6_7.

[245] Philip K. Howard. *The Death of Common Sense: How Law Is Suffocating America*. New York, NY, USA: Random House, 1994. ISBN: 978-0-8129-8274-9.

[246] Stephen E. Humphrey, Jennifer D. Nahrgang, and Frederick P. Morgeson. "Integrating motivational, social, and contextual work design features: A meta-analytic summary and theoretical extension of the work design literature". In: *Journal of Applied Psychology* 92.5 (2007), pp. 1332–1356. ISSN: 1939-1854. DOI: 10.1037/0021-9010.92.5.1332.

[247] Paul Hurley. "Consequentializing and Deontologizing: Clogging the Consequentialist Vacuum". In: *Oxford Studies in Normative Ethics*. Ed. by Mark Timmons. Vol. 3. Oxford, England, UK: Oxford University Press, 2013, pp. 123–153. ISBN: 978-0-19-968590-5. DOI: 10.1093/acprof:oso/9780199685905.003.0007.

[248] Rosalind Hursthouse. *On Virtue Ethics*. Oxford University Press, 1999. ISBN: 978-0-19-924799-8. DOI: 10.1093/0199247994.001.0001.

[249] Rosalind Hursthouse and Glen Pettigrove. "Virtue Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University, 2018. URL: https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/.

[250] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI". In: *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency*. FAccT 2021 (Virtual Event, Canada). Ed. by Madeleine Clare Elish, William Isaac, and Richard S. Zemel. New York, NY, USA: Association for Computing Machinery, 2021, pp. 624–635. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445923.

[251] Sheila Jasanoff. "Virtual, Visible, and Actionable: Data Assemblages and the Sightlines of Justice". In: *Big Data & Society* 4.2 (2017), pp. 1–15. ISSN: 2053-9517. DOI: 10.1177/2053951717724477.

[252] Agnieszka Jaworska and Julie Tannenbaum. "The Grounds of Moral Status". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2018. Metaphysics Research Lab, Stanford University, 2018. URL: https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/.

[253] Nancy S. Jecker, Caesar A. Atiure, and Martin Odei Ajei. "The Moral Standing of Social Robots: Untapped Insights from Africa". In: *Philosophy & Technology* 35.2 (2022). ISSN: 2210-5441. DOI: 10.1007/s13347-022-00531-5.

[254] Hilary Johnson and Peter Johnson. "Explanation Facilities and Interactive Systems". In: *Proceedings of the 1st International Conference on Intelligent User Interfaces*. IUI 1993 (Orlando, Florida, USA). Ed. by Wayne D. Gray, William E. Hefley, and Dianne Murray. New York, NY, USA: Association for Computing Machinery, 1993, pp. 159–166. ISBN: 978-0-89791-556-4. DOI: 10.1145/169891.169951.

[255] Robert Johnson and Adam Cureton. "Kant's Moral Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2022. Metaphysics Research Lab, Stanford University, 2022. URL: https://plato.stanford.edu/archives/fall2022/entries/kant-moral/.

[256] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. "Highly Accurate Protein Structure Prediction With AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

[257] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. "Explainable Reinforcement Learning via Reward Decomposition". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 47–53. URL: https://web.engr.oregonstate.edu/~erwig/papers/ExplainableRL_XAI19.pdf.

[258] Shelly Ian Kagan. *The Limits of Morality*. Oxford, England, UK: Oxford University Press, 1991. ISBN: 978-0-19-823916-1. DOI: 10.1093/0198239165.001.0001.

[259] Faisal Kamiran and Indrė Žliobaitė. "Explainable and Non-Explainable Discrimination in Classification". In: *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Ed. by Bart Custers, Toon Calders, Bart W. Schermer, and Tal Z. Zarsky. Studies in Applied Philosophy, Epistemology and Rational Ethics 3. Berlin/Heidelberg, Germany: Springer, 2013. Chap. 8, pp. 155–170. ISBN: 978-3-642-30487-3. DOI: 10.1007/978-3-642-30487-3_8.

[260] Immanuel Kant. *Critique of Pure Reason*. 1781.

[261] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785.

[262] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. Translated by H. J. Paton. New York, NY, USA: Harper Torchbooks, 1948 [1785].

[263] Lena Kästner, Georg Borges, Holger Hermanns, Markus Langer, Eva Schmidt, and Ulla Wessels. *Proposal for the Research Project "Explainable Intelligent Systems"*. Tech. rep. 2020.

[264] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. "On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness". In: *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops*. REW 2021 (Notre Dame, Indiana, USA). Ed. by Tao Yue and Mehdi Mirakhorli. Piscataway, NJ, USA: IEEE, 2021, pp. 169–175. ISBN: 978-1-6654-1898-0. DOI: 10.1109/REW53955.2021.00031.

[265] Leo Katz. *Ill-Gotten Gains: Evasion, Blackmail, Fraud, and Kindred Puzzles of the Law*. Chicago, IL, USA: University of Chicago Press, 1996. ISBN: 978-0-226-42593-1.

[266] Jason Kawall. "In Defense of the Primacy of the Virtues". In: *Journal of Ethics & Social Philosophy* 3.2 (2009), pp. 1–22. ISSN: 1559-3061. DOI: 10.26556/jesp.v3i2.32.

[267] Frank C. Keil. "Explanation and Understanding". In: *Annual Review of Psychology* 57.1 (2006), pp. 227–254. ISSN: 1545-2085. DOI: 10.1146/annurev.psych.57.102904.190100.

[268] Richard A. Kemmerer. "Cybersecurity". In: *Proceedings of the 25th International Conference on Software Engineering*. ICSE 2003 (Portland, Oregon, USA). Ed. by Lori A. Clarke, Laurie Dillon, and Walter F. Tichy. Piscataway, NJ, USA: IEEE, 2003, pp. 705–717. ISBN: 978-0-7695-1877-0. DOI: 10.1109/ICSE.2003.1201257.

[269] Damien Keown. "Buddhist Ethics". In: *International Encyclopedia of Ethics*. Hoboken, NJ, USA: Wiley-Blackwell, 2019, pp. 1–12. ISBN: 978-1-4443-6707-2. DOI: 10.1002/9781444367072.wbiee163.pub2.

[270] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*. NIPS 2016 (Barcelona, Spain). Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett. New York, NY, USA: Curran Associates, Inc., 2016, pp. 2280–2288. URL: https://proceedings.neurips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html.

[271] Been Kim, Cynthia Rudin, and Julie A. Shah. "The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. NIPS 2014 (Montréal, Québec, Canada). Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger. New York, NY, USA: Curran Associates, Inc., 2014, pp. 1952–1960. URL: https://proceedings.neurips.cc/paper/2014/hash/390e982518a50e280d8e2b535462ec1f-Abstract.html.

[272] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *Proceedings of the 35th International Conference on Machine Learning*. ICML 2018 (Stockholm, Sweden). Ed. by Francis Bach, Jennifer G. Dy, and Andreas Krause. Proceedings of Machine Learning Research 80. Proceedings of Machine Learning Research Press, 2018, pp. 2668–2677. URL: http://proceedings.mlr.press/v80/kim18d.html.

[273] Simon Kirchin, ed. *Reading Parfit: On What Matters*. Taylor & Francis, 2017. ISBN: 978-1-315-22553-1. DOI: 10.4324/9781315225531.

[274] René F. Kizilcec. "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface". In: *Proceedings of the 34th Conference on Human Factors in Computing Systems*. CHI 2016 (San Jose, California, USA). Ed. by Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan P. Hourcade. New York, NY, USA: Association for Computing Machinery, 2016, pp. 2390–2395. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858402.

[275] Bart P. Knijnenburg and Alfred Kobsa. "Making Decisions about Privacy: Information Disclosure in Context-Aware Recommender Systems". In: *ACM Transactions on Interactive Intelligent Systems* 3.3, 20 (2013), pp. 1–23. ISSN: 2160-6455. DOI: 10.1145/2499670.

[276] Amy J. Ko and Brad A. Myers. "Extracting and Answering Why and Why Not Questions About Java Program Output". In: *ACM Transactions on Software Engineering and Methodology* 20.2, 4 (2010), pp. 1–36. ISSN: 1049-331X. DOI: 10.1145/1824760.1824761.

[277] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. "Explainability as a Non-Functional Requirement". In: *Proceedings of the 27th IEEE International Requirements Engineering Conference*. RE 2019 (Jeju Island, South Korea). Ed. by Daniela E. Damian, Anna Perini, and Seok-Won Lee. Piscataway, NJ, USA: IEEE, 2019, pp. 363–368. ISBN: 978-1-7281-3912-8. DOI: 10.1109/RE.2019.00046.

[278] Florian Kohlbacher and Benjamin Rabe. "Leading the Way Into the Future: The Development of a (Lead) Market for Care Robotics in Japan". In: *International Journal of Technology, Policy and Management* 15.1 (2015), pp. 21–44. ISSN: 1741-5292. DOI: 10.1504/IJTPM.2015.067797.

[279] Saul A. Kripke. "Semantical Considerations on Modal Logic". In: *Acta Philosophica Fennica* 16 (1963). ISSN: 0355-1792.

[280] Maya Krishnan. "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning". In: *Philosophy & Technology* 33.3 (2020), pp. 487–502. ISSN: 2210-5441. DOI: 10.1007/s13347-019-00372-9.

[281] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. "Principles of Explanatory Debugging to Personalize Interactive Machine Learning". In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI 2015 (Atlanta, Georgia, USA). Ed. by Oliver Brdiczka, Polo Chau, Giuseppe Carenini, Shimei Pan, and Per Ola Kristensson. New York, NY, USA: Association for Computing Machinery, 2015, pp. 126–137. ISBN: 978-1-4503-3306-1. DOI: 10.1145/2678025.2701399.

[282] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models". In: *Proceedings of the 2013 IEEE Symposium on Visual Languages and Human Centric Computing*. VLHCC 2013 (San Jose, California, USA). Ed. by Caitlin Kelleher, Margaret M. Burnett, and Stefan Sauer. Piscataway, NJ, USA: IEEE, 2013, pp. 3–10. ISBN: 978-1-4799-0369-6. DOI: 10.1109/VLHCC.2013.6645235.

[283] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Amy J. Ko, and Ian Oberst. "Why-Oriented End-User Debugging of Naive Bayes Text Classification". In: *ACM Transactions on Interactive Intelligent Systems* 1.1, 2 (2011), pp. 1–31. ISSN: 2160-6463. DOI: 10.1145/2030365.2030367.

[284] P. Satheesh Kumar, M. Saravanan, and Skanda Suresh. "Explainable Classification Using Clustering in Deep Learning Models". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 115–121.

[285] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. "Automation Transparency: Implications of Uncertainty Communication for Human-Automation Interaction and Interfaces". In: *Ergonomics* 62.3 (2019), pp. 345–360. ISSN: 1366-5847. DOI: 10.1080/00140139.2018.1547842.

[286] Ray Kurzweil. *How to Create a Mind: The Secret of Human Thought Revealed*. New York, NY, USA: Viking Penguin, 2013. ISBN: 978-0-670-02529-9.

[287] Paul B. de Laat. "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" In: *Philosophy & Technology* 31.4 (2018), pp. 525–541. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0293-z.

[288] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. *An Evaluation of the Human-Interpretability of Explanation*. 2019. arXiv: 1902.00006.

[289] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. "Exploring Computational User Models for Agent Policy Summarization". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 59–65. arXiv: 1905.13271.

[290] Karel Lambert. "On Whether an Answer to a Why-Question is an Explanation if and only if it Yields Scientific Understanding". In: *Causality, Method, and Modality: Essays in Honor of Jules Vuillemin*. Ed. by Gordon G. Brittan. Dordrecht, Netherlands: Springer Netherlands, 1991, pp. 125–142. ISBN: 978-94-011-3348-7. DOI: 10.1007/978-94-011-3348-7_8.

[291] Béatrice Lamche, Ugur Adıgüzel, and Wolfgang Wörndl. "Interactive Explanations in Mobile Shopping Recommender Systems". In: *Proceedings of the Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. IntRS@RecSys 2014 (Foster City, California, USA). Ed. by Nava Tintarev, John O'Donovan, Peter Brusilovsky, Alexander Felfernig, Giovanni Semeraro, and Pasquale Lops. CEUR Workshop Proceedings 1253. CEUR-WS, 2014, pp. 14–21. URL: http://ceur-ws.org/Vol-1253/paper3.pdf.

[292] Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, and Jonas Wahl. "Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives". In: *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops*. REW 2021 (Notre Dame, Indiana, USA). Ed. by Tao Yue and Mehdi Mirakhorli. Piscataway, NJ, USA: IEEE, 2021, pp. 164–168. ISBN: 978-1-6654-1898-0. DOI: 10.1109/REW53955.2021.00030.

[293] Markus Langer, Kevin Baum, Cornelius J. König, Viviane Hähne, Daniel Oster, and Timo Speith. "Spare Me the Details: How the Type of Information About Automated Interviews Influences Applicant Reactions". In: *International Journal of Selection and Assessment* 29.2 (2021), pp. 154–169. ISSN: 1468-2389. DOI: 10.1111/ijsa.12325.

[294] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. "What Do We Want From Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research". In: *Artificial Intelligence* 296 (2021). ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103473.

[295] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. "Explainable Agency for Intelligent Autonomous Systems". In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI 2017 (San Francisco, California, USA). Ed. by Satinder P. Singh and Shaul Markovitch. Palo Alto, CA, USA: AAAI Press, 2017, pp. 4762–4764. ISBN: 978-1-57735-780-3. URL: http://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15046.

[296] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "Analyzing Classifiers: Fisher Vectors and Deep Neural Networks". In: *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR 2016 (Las Vegas, Nevada, USA). Ed. by Tinne Tuytelaars, Fei-Fei Li, Ruzena Bajcsy, Lourdes Agapito, Tamara Berg, Jana Kosecka, and Lihi Zelnik-Manor. Piscataway, NJ, USA: IEEE, 2016, pp. 2912–2920. ISBN: 978-1-4673-8850-4. DOI: 10.1109/CVPR.2016.318.

[297] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn". In: *Nature Communications* 10.1, 1096 (2019), pp. 1–8. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08987-4.

[298] Averill M. Law and W. David Kelton. *Simulation Modeling and Analysis*. Third Edition. New York, NY, USA: McGraw-Hill Higher Education, 2007. ISBN: 978-0-07-059292-6.

[299] George Lawton. *UX Defines Chasm Between Explainable vs. Interpretable AI*. 2019. URL: https://searchenterpriseai.techtarget.com/feature/UX-defines-chasm-between-explainable-vs-interpretable-AI (visited on 01/30/2022).

[300] Derek Leben. "A Rawlsian Algorithm for Autonomous Vehicles". In: *Ethics and Information Technology* 19.2 (2017), pp. 107–115. ISSN: 1572-8439. DOI: 10.1007/s10676-017-9419-3.

[301] John D. Lee and Katrina A. See. "Trust in Automation: Designing for Appropriate Reliance". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (2004), pp. 50–80. ISSN: 1547-8181. DOI: 10.1518/hfes.46.1.50_30392.

[302] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. "Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation". In: *Proceedings of the 2019 ACM on Human-Computer Interaction* 3.CSCW, 182 (2019), pp. 1–26. ISSN: 2573-0142. DOI: 10.1145/3359284.

[303] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes". In: *Philosophy & Technology* 31.4 (2018), pp. 611–627. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0279-x.

[304] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences". In: *Proceedings of the 38th Conference on Human Factors in Computing Systems*. CHI 2020 (Honolulu, Hawaii, USA). Ed. by Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane P. Samson, and Rafal Kocielnik. New York, NY, USA: ACM, 2020, pp. 1–15. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376590.

[305] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. "*Why and Why Not* Explanations Improve the Intelligibility of Context-Aware Intelligent Systems". In: *Proceedings of the 27th Conference on Human Factors in Computing Systems*. CHI 2009 (Boston, Massachusetts, USA). Ed. by Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith R. Morris, Scott E. Hudson, and Saul Greenberg. New York, NY, USA: Association for Computing Machinery, 2009, pp. 2119–2128. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1519023.

[306] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *Commununications of the ACM* 61.10 (2018), pp. 36–43. ISSN: 0001-0782. DOI: 10.1145/3233231.

[307] Huafeng Liu, Jingxuan Wen, Liping Jing, Jian Yu, Xiangliang Zhang, and Min Zhang. "In2Rec: Influence-Based Interpretable Recommendation". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM 2019 (Beijing, China). Ed. by Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1803–1812. ISBN: 978-1-4503-6976-3. DOI: 10.1145/3357384.3358017.

[308] Tania Lombrozo. "The Instrumental Value of Explanations". In: *Philosophy Compass* 6.8 (2011), pp. 539–551. ISSN: 1747-9991. DOI: 10.1111/j.1747-9991.2011.00413.x.

[309] Errol Lord and Barry Maguire. *Weighing Reasons*. Oxford, England, USA: Oxford University Press, 2016. ISBN: 978-0-19-931519-2. DOI: 10.1093/acprof:oso/9780199315192.001.0001.

[310] Jacobus Lubsen, J. Pool, and E." Van der Does. "A Practical Device for the Application of a Diagnostic or Prognostic Function". In: *Methods of Information in Medicine* 17.2 (1978), pp. 127–129. ISSN: 0026-1270. DOI: 10.1055/s-0038-1636613.

[311] Crisrael Lucero, Braulio Coronado, Oliver Hui, and Douglas S. Lange. "Exploring Explainable Artificial Intelligence and Autonomy Through Provenance". In: *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. IJCAI/ECAI XAI 2018 (Stockholm, Sweden). Ed. by David W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni. 2018, pp. 85–89.

[312] Ana Lucic, Hinda Haned, and de Rijke Maarten. "Contrastive Explanations for Large Errors in Retail Forecasting Predictions through Monte Carlo Simulations". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 66–72. URL: https://a-lucic.github.io/talks/ijcai2019-ana.pdf.

[313] Christoph Luetge. "The German Ethics Code for Automated and Connected Driving". In: *Philosophy & Technology* 30 (2017), pp. 547–558. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0284-0.

[314] Peter Machamer, Lindley Darden, and Carl F. Craver. "Thinking About Mechanisms". In: *Philosophy of Science* 67.1 (2000), pp. 1–25. ISSN: 1539-767X. DOI: 10.1086/392759.

[315] John L. Mackie. "The Disutility of Act-Utilitarianism". In: *The Philosophical Quarterly* 23.93 (1973), pp. 289–300. ISSN: 1467-9213. DOI: 10.2307/2218058.

[316] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. "Explainable Reinforcement Learning Through a Causal Lens". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 73–79. arXiv: 1905.10958.

[317] Sara Mann, Barnaby Crook, Lena Kästner, Astrid Schomäcker, and Timo Speith. "Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy". In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*. REW 2023 (Hannover, Germany). Ed. by Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider. Piscataway, NJ, USA: IEEE, 2023.

[318] Susanne Mantel. "Worldly Reasons: An Ontological Inquiry Into Motivating Considerations and Normative Reasons". In: *Pacific Philosophical Quarterly* 98.S1 (2017), pp. 5–28. ISSN: 0279-0750. DOI: 10.1111/papq.12094.

[319] Susanne Mantel. *Determined by Reasons: A Competence Account of Acting for a Normative Reason*. New York, NY, USA: Routledge, 2018. ISBN: 978-1-351-18635-3. DOI: 10.4324/9781351186353.

[320] David Marr. *Vision: A Computational Investigation into the Human Representation of Visual Information*. San Francisco, CA, USA: W.H. Freeman & Company, 1982. ISBN: 978-0-7167-1284-8.

[321] David Marr and Tomaso Poggio. "From Understanding Computation to Understanding Neural Circuitry". In: *AI Memos* (1976). AIM-357. URL: http://hdl.handle.net/1721.1/5782.

[322] Sherin Mary Mathews. "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review". In: *Intelligent Computing – Proceedings of the 2019 Computing Conference*. CompCom 2019 (London, England, UK). Ed. by Kohei Arai, Rahul Bhatia, and Supriya Kapoor. Advances in Intelligent Systems and Computing 998. Cham, Switzerland: Springer International Publishing, 2019, pp. 1269–1292. ISBN: 978-3-030-22868-2. DOI: 10.1007/978-3-030-22868-2_90.

[323] Andreas Matthias. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata". In: *Ethics and Information Technology* 6.3 (2004), pp. 175–183. ISSN: 1572-8439. DOI: 10.1007/s10676-004-3422-1.

[324] G. Randolph Mayes. "Theories of Explanation". In: *The Internet Encyclopedia of Philosophy* (2021). ISSN: 2161-0002. URL: https://www.iep.utm.edu/explanat/ (visited on 11/01/2021).

[325] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. "Thinking Positively – Explanatory Feedback for Conversational Recommender Systems". In: *Proceedings of the 7th European Conference on Case-Based Reasoning Explanation Workshop*. ECCBR WS 2004 (Madrid, Spain). 2004, pp. 115–124.

[326] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. "Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys 2018 (Vancouver, British Columbia, Canada). Ed. by Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan. New York, NY, USA: Association for Computing Machinery, 2018, pp. 31–39. ISBN: 978-1-4503-5901-6. DOI: 10.1145/3240323.3240354.

[327] Bruce M. McLaren. "Extensionally Defining Principles and Cases in Ethics: An AI Model". In: *Artificial Intelligence* 150.1 (2003). Special Issue on AI and Law, pp. 145–181. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(03)00135-8.

[328] Bruce M. McLaren. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 29–37. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.67.

[329] Bruce M. McLaren. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions". In: *Machine Ethics*. Ed. by Michael Anderson and Susan Leigh Anderson. New York, NY, USA: Cambridge University Press, 2011, pp. 297–315. ISBN: 978-1-108-46175-7.

[330] Bruce M. McLaren and Kevin D Ashley. "Case-Based Comparative Evaluation in TRUTH-TELLER". In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Ed. by Johanna D. Moore and Jill Fain Lehman. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1995, pp. 72–77. ISBN: 978-0-8058-2159-8.

[331] Carolyn McLeod. "Trust". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/fall2020/entries/trust/.

[332] Paul McNamara and Frederik Van De Putte. "Deontic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022. URL: https://plato.stanford.edu/archives/fall2022/entries/logic-deontic/.

[333] Hugo Mercier. "The Argumentative Theory: Predictions and Empirical Evidence". In: *Trends in Cognitive Sciences* 20.9 (2016), pp. 689–700. ISSN: 1364-6613. DOI: 10.1016/j.tics.2016.07.001.

[334] Hugo Mercier and Dan Sperber. "Why Do Humans Reason? Arguments for an Argumentative Theory". In: *Behavioral and Brain Sciences* 34 (2011), pp. 57–111. ISSN: 1469-1825. DOI: 10.1017/S0140525X10000968.

[335] Rebecca T. Mercuri and Peter G. Neumann. "Security by Obscurity". In: *Communications of the ACM* 46.11 (2003), p. 160. ISSN: 1557-7317. DOI: 10.1145/948383.948413.

[336] Merriam-Webster Dictionary. *Stakeholder*. 2022. URL: https://www.merriam-webster.com/dictionary/stakeholder (visited on 01/30/2022).

[337] Loizos Michael. "Machine Coaching". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 80–86. DOI: 10.5281/zenodo.3931266.

[338] John Stuart Mill. *Utilitarianism*. 1861.

[339] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702. DOI: 10.1016/j.artint.2018.07.007.

[340] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Australia). Ed. by David W. Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. Santa Clara County, CA, USA: IJCAI, 2017, pp. 36–42. arXiv: 1712.00547.

[341] Brent D. Mittelstadt, Chris Russell, and Sandra Wachter. "Explaining Explanations in AI". In: *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*. FAT* 2019. Ed. by Danah Boyd and Jamie H. Morgenstern. Atlanta, Georgia, USA: Association for Computing Machinery, 2019, pp. 279–288. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287574.

[342] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The Ethics of Algorithms: Mapping the Debate". In: *Big Data & Society* 3.2 (2016), pp. 1–21. ISSN: 2053-9517. DOI: 10.1177/2053951716679679.

[343] Christoph Molnar. *Interpretable Machine Learning – A Guide for Making Black Box Models Explainable*. Victoria, British Columbia, Canada: Leanpub, 2019. URL: https://christophm.github.io/interpretable-ml-book/.

[344] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Methods for Interpreting and Understanding Deep Neural Networks". In: *Digital Signal Processing* 73 (2018), pp. 1–15. ISSN: 1051-2004. DOI: 10.1016/j.dsp.2017.10.011.

[345] Isaac Monteath and Raymond Sheh. "Assisted and Incremental Medical Diagnosis Using Explainable Artificial Intelligence". In: *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. IJCAI/ECAI XAI 2018 (Stockholm, Sweden). Ed. by David W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni. 2018, pp. 104–108.

[346] James H. Moor. "The Nature, Importance, and Difficulty of Machine Ethics". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 18–21. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.80.

[347] Martin Možina, Janez Demšar, Michael Kattan, and Blaž Zupan. "Nomograms for Visualization of Naive Bayesian Classifier". In: *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*. PKDD 2004 (Pisa, Italy). Ed. by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi. Lecture Notes in Computer Science 3202. Berlin/Heidelberg, Germany: Springer, 2004, pp. 337–348. ISBN: 978-3-540-30116-5. DOI: 10.1007/978-3-540-30116-5_32.

[348] Sayooran Nagulendra and Julita Vassileva. "Providing Awareness, Explanation and Control of Personalized Filtering in a Social Networking Site". In: *Information Systems Frontiers* 18.1 (2016), pp. 145–158. DOI: 10.1007/s10796-015-9577-y.

[349] Grzegorz J. Nalepa, Martijn van Otterlo, Szymon Bobek, and Martin Atzmueller. "From Context Mediation to Declarative Values and Explainability". In: *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. IJCAI/ECAI XAI 2018 (Stockholm, Sweden). Ed. by David W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni. 2018, pp. 109–113.

[350] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. *How Do Humans Understand Explanations From Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation*. 2018. arXiv: 1802.00682.

[351] Allen Newell. "The Knowledge Level". In: *Artificial Intelligence* 18.1 (1982), pp. 87–127. ISSN: 0004-3702. DOI: 10.1016/0004-3702(82)90012-1.

[352] Allen Newell. "The Intentional Stance and the Knowledge Level". In: *Behavioral and Brain Sciences* 11.3 (1988), pp. 520–522. ISSN: 1469-1825. DOI: 10.1017/S0140525X00058763.

[353] Hoa Nguyen. *Guided Backpropagation with TF2*. 2022. URL: https://colab.research.google.com/drive/17tAC7xx2IJxjK700bdaLatTVeDA02GJn (visited on 10/31/2022).

[354] Merel Noorman. "Computing and Moral Responsibility". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

[355] Ulrich Nortmann. *Deontische Logik ohne Paradoxien: Semantik und Logik des Normativen*. München, Germany: Philosophia, 1989. ISBN: 978-3-88405-067-5.

[356] Ulrich Nortmann and Timo Speith. *Skript: Einführung in die Wissenschaftstheorie*. Lecture notes. Universität des Saarlandes, 2018.

[357] Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker. "The Impact of Explanation Dialogues on Human-Computer Trust". In: *Proceedings of the 15th International Conference on Human-Computer Interaction*. HCI 2013. Ed. by Masaaki Kurosu. Lecture Notes in Computer Science 8006. Part III. Las Vegas, Nevada, USA: Springer, 2013, pp. 59–67. ISBN: 978-3-642-39265-8. DOI: 10.1007/978-3-642-39265-8_7.

[358] Ingrid Nunes and Dietmar Jannach. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems". In: *User Modeling and User-Adapted Interaction* 27.3–5 (2017), pp. 393–444. ISSN: 1573-1391. DOI: 10.1007/s11257-017-9195-0.

[359] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Broadway Books, 2016. ISBN: 978-0-553-41881-1.

[360] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization". In: *Distill* (2017). ISSN: 2476-0757. DOI: 10.23915/distill.00007.

[361] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. "The Building Blocks of Interpretability". In: *Distill* (2018). DOI: 10.23915/distill.00010.

[362] Matthew L. Olson, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. "Counterfactual States for Atari Agents via Generative Deep Learning". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 87–93. arXiv: 1909.12969.

[363] Daniel Oster. "Explanation Requirements Concerning Artificial Systems – How to Transmit Transparency From Designer Side to Non-Designer Side". MA thesis. Saarland University, 2019.

[364] James A. Overton. "Explanation in Science". PhD thesis. The University of Western Ontario, 2012.

[365] Andrés Páez. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)". In: *Minds and Machines* 29.3 (2019), pp. 441–459. DOI: 10.1007/s11023-019-09502-w.

[366] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. "How Model Accuracy and Explanation Fidelity Influence User Trust in AI". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 94–100. arXiv: 1907.12652.

[367] Raja Parasuraman and Dietrich H. Manzey. "Complacency and Bias in Human Use of Automation: An Attentional Integration". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52.3 (2010), pp. 381–410. ISSN: 1547-8181. DOI: 10.1177/0018720810376055.

[368] Raja Parasuraman and Victor Riley. "Humans and Automation: Use, Misuse, Disuse, Abuse". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39.2 (1997), pp. 230–253. ISSN: 1547-8181. DOI: 10.1518/001872097778543886.

[369] Derek Parfit. *On What Matters: Volume One*. Oxford, England, UK: Oxford University Press, 2011. ISBN: 978-0-19-957280-9. DOI: 10.1093/acprof:osobl/9780199572809.001.0001.

[370] Derek Parfit. *On What Matters: Volume Two*. Oxford, England, UK: Oxford University Press, 2011. ISBN: 978-0-19-957281-6. DOI: 10.1093/acprof:osobl/9780199572816.001.0001.

[371] Derek Parfit. *On What Matters: Volume Three*. Oxford, England, UK: Oxford University Press, 2017. ISBN: 978-0-19-877860-8. DOI: 10.1093/oso/9780198778608.001.0001.

[372] Régis Pierrard, Jean-Philippe Poli, and Céline Hudelot. "A New Approach for Explainable Multiple Organ Annotation with Few Data". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 101–107. arXiv: 1912.12932.

[373] Wolter Pieters. "Explanation and Trust: What to Tell the User in Security and AI?" In: *Ethics and Information Technology* 13.1 (2011), pp. 53–64. ISSN: 1572-8439. DOI: 10.1007/s10676-010-9253-3.

[374] Matthijs A. Pontier and Johan F. Hoorn. "Toward Machines That Behave Ethically Better Than Humans Do". In: *Proceedings of of the 34th International Annual Conference of the Cognitive Science Society*. CogSci 2012 (Sapporo, Japan). Ed. by Naomi Miyake, David Peebles, and Richard P. Cooper. Austin, Texas, USA: Cognitive Science Society, 2012, pp. 2198–2203. ISBN: 978-1-62276-304-7. URL: https://escholarship.org/uc/item/4rj8883h.

[375] Douglas W. Portmore. "Consequentializing Moral Theories". In: *Pacific Philosophical Quarterly* 88.1 (2007), pp. 39–73. ISSN: 0279-0750. DOI: 10.1111/j.1468-0114.2007.00280.x.

[376] Douglas W. Portmore. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford, England, UK: Oxford University Press, 2011. ISBN: 978-0-19-979453-9. DOI: 10.1093/acprof:oso/9780199794539.001.0001.

[377] Thomas M. Powers. "Prospects for a Kantian Machine". In: *IEEE Intelligent Systems* 21.4 (2006), pp. 46–51. ISSN: 1541-1672. DOI: 10.1109/MIS.2006.77.

[378] Thomas M. Powers. "Incremental Machine Ethics". In: *IEEE Robotics & Automation Magazine* 18.1 (2011), pp. 51–58. ISSN: 1558-223X. DOI: 10.1109/MRA.2010.940152.

[379] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. *Stakeholders in Explainable AI*. arXiv: 1810.00184.

[380] Arthur N. Prior. *Past, Present and Future*. Vol. 154. Oxford, England, UK: Oxford University Press, 1967. ISBN: 978-0-19-824311-3. DOI: 10.1093/acprof:oso/9780198243113.001.0001.

[381] Stathis Psillos. *Causation and Explanation*. Abingdon, England, UK: Routledge, 2002. ISBN: 978-1-902683-42-3.

[382] Pearl Pu and Li Chen. "Trust Building with Explanation Interfaces". In: *Proceedings of the 11th International Conference on Intelligent User Interfaces*. IUI 2006 (Sydney, New South Wales, Australia). Ed. by Cécile Paris and Candace L. Sidner. New York, NY, USA: Association for Computing Machinery, 2006, pp. 93–100. ISBN: 978-1-59593-287-7. DOI: 10.1145/1111449.1111475.

[383] Pearl Pu and Li Chen. "Trust-Inspiring Explanation Interfaces for Recommender Systems". In: *Knowledge-Based Systems* 20.6 (2007), pp. 542–556. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2007.04.004.

[384] Sharon L. S. Purkiss, Pamela L. Perrewé, Treena L. Gillespie, Bronston T. Mayes, and Gerald R. Ferris. "Implicit Sources of Bias in Employment Interview Judgments and Decisions". In: *Organizational Behavior and Human Decision Processes* 101.2 (2006), pp. 152–167. DOI: 10.1016/j.obhdp.2006.06.005.

[385] Vanessa Putnam and Cristina Conati. "Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS)". In: *Joint Proceedings of the 24th ACM Conference on Intelligent User Interfaces Workshops*. IUI WS 2019 (Los Angeles, California, USA). Ed. by Christoph Trattner, Denis Parra, and Nathalie Riche. CEUR Workshop Proceedings 2327. CEUR-WS, 2019, pp. 1–7. URL: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-19.pdf.

[386] Vanessa Putnam, Lea Rieger, and Cristina Conati. "Towards Personalized XAI: A Case Study in Intelligent Tutoring Systems". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 108–114. arXiv: 1912.04464.

[387] Zenon W. Pylyshyn. "Computing in Cognitive Science". In: *Foundations of Cognitive Science*. Ed. by Michael I. Posner. Cambridge, MA, USA: The MIT Press, 1989, pp. 51–91. ISBN: 978-0-262-28180-5.

[388] Emilee J. Rader, Kelley Cotter, and Janghee Cho. "Explanations as Mechanisms for Supporting Algorithmic Transparency". In: *Proceedings of the 36th Conference on Human Factors in Computing Systems*. CHI 2018 (Montréal, Québec, Canada). Ed. by Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox. New York, NY, USA: Association for Computing Machinery, 2018, 103, pp. 1–13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173677.

[389] Nazneen F. Rajani and Raymond J. Mooney. "Using Explanations to Improve Ensembling of Visual Question Answering Systems". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 43–47.

[390] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing". In: *Proceedings of the 3rd ACM Conference on Fairness, Accountability, and Transparency*. FAT* 2020 (Barcelona, Spain). Ed. by Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna. New York, NY, USA: Association for Computing Machinery, 2020, pp. 33–44. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372873.

[391] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. “Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges”. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Ed. by Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven. Cham, Switzerland: Springer International Publishing, 2018. Chap. 2, pp. 19–36. ISBN: 978-3-319-98131-4. DOI: 10.1007/978-3-319-98131-4_2.

[392] John Rawls. “Outline of a Decision Procedure for Ethics”. In: *The Philosophical Review* 60.2 (1951), pp. 177–197. ISSN: 0031-8108. DOI: 10.2307/2181696.

[393] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 2016 (San Francisco, California, USA). Ed. by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

[394] Ariella Richardson and Avi Rosenfeld. “A Survey of Interpretability and Explainability in Human-Agent Systems”. In: *Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. IJCAI/ECAI XAI 2018 (Stockholm, Sweden). Ed. by David W. Aha, Trevor Darrell, Patrick Doherty, and Daniele Magazzeni. 2018, pp. 137–143.

[395] Mark O. Riedl. “Human-Centered Artificial Intelligence and Machine Learning”. In: *Human Behavior and Emerging Technologies* 1.1 (2019), pp. 33–36. ISSN: 2578-1863. DOI: 10.1002/hbe2.117.

[396] Ludovic Righetti, Q.-C. Pham, Raj Madhavan, and Raja Chatila. “Lethal Autonomous Weapon Systems [Ethical, Legal, and Societal Issues]”. In: *IEEE Robotics & Automation Magazine* 25.1 (2018), pp. 123–126. ISSN: 1558-223X. DOI: 10.1109/MRA.2017.2787267.

[397] Scott Robbins. “A Misdirected Principle with a Catch: Explicability for AI”. In: *Minds and Machines* 29.4 (2019), pp. 495–514. ISSN: 1572-8641. DOI: 10.1007/s11023-019-09509-3.

[398] Michael Rohlf. “Immanuel Kant”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/fall2020/entries/kant/.

[399] Avi Rosenfeld and Ariella Richardson. “Explainability in Human–Agent Systems”. In: *Autonomous Agents and Multi-Agent Systems* 33.6 (2019), pp. 673–705. ISSN: 1573-7454. DOI: 10.1007/s10458-019-09408-y.

[400] Stephanie Rosenthal, Sai P. Selvaraj, and Manuela M. Veloso. “Verbalization: Narration of Autonomous Robot Experience”. In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. IJCAI 2016 (New York City, New York, USA). Ed. by Subbarao Kambhampati. Palo Alto, CA, USA: IJCAI/AAAI Press, 2016, pp. 862–868. ISBN: 978-1-57735-770-4. URL: http://www.ijcai.org/Abstract/16/127.

[401] Alf Ross. “Imperatives and Logic”. In: *Philosophy of Science* 11.1 (1944), pp. 30–46. ISSN: 1539-767X. DOI: 10.1086/286823.

[402] William D. Ross. *The Right and the Good*. Oxford, England, UK: Oxford University Press, 1930.

[403] William D. Ross. *The Right and the Good (Second Edition)*. Oxford, England, UK: Oxford University Press, 2002. ISBN: 978-0-19-925265-7. DOI: 10.1093/0199252653.001.0001.

[404] Cynthia Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[405] Fernando Rudy-Hiller. “The Epistemic Condition for Moral Responsibility”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2018. Metaphysics Research Lab, Stanford University, 2018. URL: https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/.

[406] Wesley C. Salmon. “Statistical Explanation”. In: *Statistical Explanation and Statistical Relevance*. Ed. by Wesley C. Salmon. Pittsburgh, PA, USA: University of Pittsburgh Press, 1971. Chap. 3, pp. 29–87. ISBN: 978-0-8229-5225-1. URL: https://upittpress.org/books/9780822952251/.

[407] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ, USA: Princeton University Press, 1984. ISBN: 978-0-691-22148-9. DOI: 10.1515/9780691221489.

[408] Wesley C. Salmon. *Causality and Explanation*. Oxford, England, UK: Oxford University Press, 1998. ISBN: 978-0-19-510864-4. DOI: 10.1093/0195108647.001.0001.

[409] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. 2017. arXiv: 1708.08296.

[410] Masahiro Sato, Koki Nagatani, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. "Context Style Explanation for Recommender Systems". In: *Journal of Information Processing* 27 (2019), pp. 720–729. ISSN: 0387-6101. DOI: 10.2197/ipsjjip.27.720.

[411] Neil Savage. "Robots Rise to Meet the Challenge of Caring for Old People". In: *Nature* 601.7893 (2022), pp. 8–10. ISSN: 1476-4687. DOI: 10.1038/d41586-022-00072-z.

[412] Pete Sawyer, Nelly Bencomo, Jon Whittle, Emmanuel Letier, and Anthony Finkelstein. "Requirements-Aware Systems: A Research Agenda for RE for Self-Adaptive Systems". In: *Proceedings of the 18th IEEE International Requirements Engineering Conference*. RE 2010 (Sydney, New South Wales, Australia). Ed. by Jane Cleland-Huang and Didar Zowghi. Piscataway, NJ, USA: IEEE, 2010, pp. 95–103. ISBN: 978-0-7695-4162-4. DOI: 10.1109/RE.2010.21.

[413] Geoff Sayre-McCord. "Metaethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2014. Metaphysics Research Lab, Stanford University, 2014. URL: https://plato.stanford.edu/archives/sum2014/entries/metaethics/.

[414] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek F. Abdelzaher, and John O'Donovan. "Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis". In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI 2015 (Atlanta, Georgia, USA). Ed. by Oliver Brdiczka, Polo Chau, Giuseppe Carenini, Shimei Pan, and Per Ola Kristensson. New York, NY, USA: Association for Computing Machinery, 2015, pp. 345–356. ISBN: 978-1-4503-3306-1. DOI: 10.1145/2678025.2701406.

[415] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. "I Can Do Better Than Your AI: Expertise and Explanations". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 240–251. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302308.

[416] Eva Schmidt. "Normative Reasons for Mentalism". In: *Metaepistemology: Realism and Anti-Realism*. Ed. by Christos Kyriacou and Robin McKenna. Cham, Switzerland: Springer International Publishing, 2018. Chap. 5, pp. 97–120. ISBN: 978-3-319-93369-6. DOI: 10.1007/978-3-319-93369-6_5.

[417] Johannes Schneider and Joshua Peter Handali. "Personalized Explanation for Machine Learning: A Conceptualization". In: *Proceedings of the 27th European Conference on Information Systems*. ECIS 2019 (Stockholm and Uppsala, Sweden). Ed. by Jan vom Brocke, Shirley Gregor, and Oliver Müller. 2019. ISBN: 978-1-7336325-0-8. URL: https://aisel.aisnet.org/ecis2019_rp/171.

[418] Arthur Schopenhauer. *On the Basis of Morality*. 1840.

[419] Gerhard Schurz. "Wissenschaftliche Erklärung". In: *Wissenschaftstheorie*. Ed. by Andreas Bartels and Manfred Stöckler. Paderborn, Germany: Mentis, 2009, pp. 69–88. ISBN: 978-3-89785-591-5.

[420] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the 16th IEEE International Conference on Computer Vision*. ICCV 2017 (Venice, Italy). Ed. by Katsushi Ikeuchi, Gérard Medioni, Marcello Pelillo, Rita Cucchiara, Yasuyuki Matsushita, Nicu Sebe, and Stefano Soatto. Piscataway, NJ, USA: IEEE, 2017, pp. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74.

[421] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7.

[422] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A. Keim, and Mennatallah El-Assady. "Going Beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models". In: *Proceedings of the 1st Workshop on Visualization for AI Explainability*. VISxAI@IEEE VIS 2018 (Berlin, Germany). Ed. by Mennatallah El-Assady, Duen Horng (Polo) Chau, Adam Perer, Hendrik Strobelt, and Fernanda Viégas. 2018. URL: https://kops.uni-konstanz.de/handle/123456789/45045.

[423] Daniel B. Shank, Alyssa DeSanti, and Timothy Maninger. "When Are Artificial Intelligence Versus Human Agents Faulted for Wrongdoing? Moral Attributions After Individual and Joint Decisions". In: *Information, Communication & Society* 22.5 (2019), pp. 648–663. ISSN: 1369-118X. DOI: 10.1080/1369118X.2019.1568515.

[424] Amanda Sharkey and Noel Sharkey. "Granny and the Robots: Ethical Issues in Robot Care for the Elderly". In: *Ethics and Information Technology* 14.1 (2012), pp. 27–40. ISSN: 1572-8439. DOI: 10.1007/s10676-010-9234-6.

[425] Raymond Ka-Man Sheh. "Different XAI for Different HRI". In: *Proceedings of the 2017 AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction* (Arlington, Virginia, USA). Ed. by Laura Hiatt and Elin A. Topp. AAAI Technical Report FS-17-01. Palo Alto, CA, USA: AAAI Press, 2017, pp. 114–117. ISBN: 978-1-57735-794-0.

[426] Raymond Ka-Man Sheh. "'Why did you do that?' Explainable Intelligent Robots". In: *Workshops of the 31st AAAI Conference on Artificial Intelligence* (San Francisco, California, USA). Ed. by Kartik Talamadupula, Shirin Sohrabi, Loizos Michael, and Biplav Srivastava. AAAI Technical Report WS-17-10. Palo Alto, CA, USA: AAAI Press, 2017, pp. 628–634. ISBN: 978-1-57735-786-5.

[427] Raymond Ka-Man Sheh and Isaac Monteath. "Defining Explainable AI for Requirements Analysis". In: *KI – Künstliche Intelligenz* 32.4 (2018), pp. 261–266. ISSN: 1610-1987. DOI: 10.1007/s13218-018-0559-3.

[428] Edward H. Shortliffe and Bruce G. Buchanan. "A Model of Inexact Reasoning in Medicine". In: *Mathematical Biosciences* 23.3-4 (1975), pp. 351–379. ISSN: 0025-5564. DOI: 10.1016/0025-5564(75)90047-4.

[429] May Sim. "Confucian and Daoist Virtue Ethics". In: *Varieties of Virtue Ethics*. Ed. by David Carr, James Arthur, and Kristján Kristjánsson. London, England, UK: Palgrave Macmillan, 2017, pp. 105–121. ISBN: 978-1-137-59177-7. DOI: 10.1057/978-1-137-59177-7_7.

[430] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2013. arXiv: 1312.6034.

[431] Peter Singer. "Is Act-Utilitarianism Self-Defeating?" In: *The Philosophical Review* 81.1 (1972), pp. 94–104. ISSN: 1558-1470. DOI: 10.2307/2184228.

[432] Peter Singer. *Practical Ethics*. Third Edition; Original Published in 1979. Cambridge, England, UK: Cambridge University Press, 2011. ISBN: 978-0-521-70768-8.

[433] Peter Singer. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Revised Edition; Original Published in 1981. Princeton, NJ, USA: Princeton University Press, 2011. ISBN: 978-0-691-15069-7.

[434] Rashmi Sinha and Kirsten Swearingen. "The Role of Transparency in Recommender Systems". In: *Extended Abstracts of the 20th Conference on Human Factors in Computing Systems*. CHI EA 2002 (Minneapolis, Minnesota, USA). Ed. by Loren G. Terveen and Dennis R. Wixon. New York, Ny, USA: Association for Computing Machinery, 2002, pp. 830–831. ISBN: 1-58113-454-1. DOI: 10.1145/506443.506619.

[435] Walter Sinnott-Armstrong. "Consequentialism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2021. Metaphysics Research Lab, Stanford University, 2021. URL: https://plato.stanford.edu/archives/fall2021/entries/consequentialism/.

[436] Elizabeth I. Sklar and Mohammad Q. Azhar. "Explanation Through Argumentation". In: *Proceedings of the 6th International Conference on Human-Agent Interaction*. HAI 2018 (Southampton, England, UK). Ed. by Michita Imai, Tim Norman, Elizabeth Sklar, and Takanori Komatsu. New York, NY, USA: Association for Computing Machinery, 2018, pp. 277–285. ISBN: 978-1-4503-5953-5. DOI: 10.1145/3284432.3284470.

[437] John J. C. Smart. "Can Biology Be an Exact Science?" In: *Synthese* 11.4 (1959), pp. 359–368. ISSN: 1573-0964. DOI: 10.1007/BF00486197.

[438] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. "SmoothGrad: Removing Noise by Adding Noise". In: *Proceedings of the International Conference on Machine Learning Workshop on Visualization for Deep Learning*. ICML VIS 2017 (Sydney, New South Wales, Australia). Ed. by John Canny, Polo Chau, Xiangmin Fan, Biye Jiang, and Jun-Yan Zhu. 2017. arXiv: 1706.03825.

[439] Joshua K. Smith. *Robotic Persons: Our Future With Social Robots*. Bloomington, IN, USA: Westbow Press, 2021. ISBN: 978-1-6642-1974-8.

[440] Michael Smith. "The Humean Theory of Motivation". In: *Mind* 96.381 (1987), pp. 36–61. ISSN: 0026-4423. DOI: 10.1093/mind/XCVI.381.36.

[441] Kacper Sokol and Peter A. Flach. "Conversational Explanations of Machine Learning Predictions Through Class-Contrastive Counterfactual Statements". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI 2018. Ed. by Jérôme Lang. IJCAI Organization, 2018, pp. 5785–5786. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/836.

[442] Kacper Sokol and Peter A. Flach. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* 2020 (Barcelona, Spain). Ed. by Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna. New York, NY, USA: Association for Computing Machinery, 2020, pp. 56–67. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372870.

[443] Kacper Sokol and Peter A. Flach. "One Explanation Does Not Fit All". In: *KI – Künstliche Intelligenz* 34.2 (2020), pp. 235–250. ISSN: 1610-1987. DOI: 10.1007/s13218-020-00637-y.

[444] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. "Explanation in Case-Based Reasoning – Perspectives and Goals". In: *Artificial Intelligence Review* 24.2 (2005), pp. 109–143. ISSN: 1573-7462. DOI: 10.1007/s10462-005-4607-7.

[445] Robert Sparrow. "The Turing Triage Test". In: *Ethics and Information Technology* 6.4 (2004), pp. 203–213. ISSN: 1572-8439. DOI: 10.1007/s10676-004-6491-2.

[446] Timo Speith. "Towards a Framework of Verifiable Machine Ethics and Machine Explainability". MA thesis. Saarland University, 2018.

[447] Timo Speith. "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods". In: *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency*. FAccT 2022 (Seoul, Republic of Korea). Ed. by Charles Isbell, Seth Lazar, Alice Oh, and Alice Xiang. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2239–2250. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3534639.

[448] Timo Speith. "How to Evaluate Explainability – A Case for Three Criteria". In: *Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops*. REW 2022 (Virtual Event, Australia). Ed. by Eric Knauss, Gunter Mussbacher, Chetan Arora, Muneera Bano, and Jean-Guy Schneider. Piscataway, NJ, USA: IEEE, 2022, pp. 92–97. ISBN: 978-1-6654-6000-2. DOI: 10.1109/REW56159.2022.00024.

[449] Timo Speith and Markus Langer. "A New Perspective on Evaluation Methods for Explainable Artificial Intelligence (XAI)". In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*. REW 2023 (Hannover, Germany). Ed. by Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider. Piscataway, NJ, USA: IEEE, 2023.

[450] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. "Striving for Simplicity: The All Convolutional Net". In: *Proceedings of the 3rd International Conference on Learning Representations Workshop Track*. ICLR WT 2015 (San Diego, California, USA). Ed. by Yoshua Bengio and Yann LeCun. 2015. arXiv: 1412.6806.

[451] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. "Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation". In: *Proceedings of the 28th International Conference on Automated Planning and Scheduling*. ICAPS 2018 (Delft, The Netherlands). Ed. by Mathijs de Weerdt, Sven Koenig, Gabriele Röger, and Matthijs T. J. Spaan. Palo Alto, CA, USA: AAAI Press, 2018, pp. 518–526. ISBN: 978-1-57735-797-1. URL: https://ojs.aaai.org/index.php/ICAPS/article/view/13930.

[452] Mohan Sridharan and Ben Meadows. "Towards a Theory of Explanations for Human–Robot Collaboration". In: *KI – Künstliche Intelligenz* 33.4 (2019), pp. 331–342. ISSN: 1573-7462. DOI: 10.1007/s13218-019-00616-y.

[453] Michael Stocker. "The Schizophrenia of Modern Ethical Theories". In: *The Journal of Philosophy* 73.14 (1977), pp. 453–466. ISSN: 1939-8549. DOI: 10.2307/2025782.

[454] Michael T. Stuart and Nancy J. Nersessian. "Peeking Inside the Black Box: A New Kind of Scientific Visualization". In: *Minds and Machines* 29.1 (2019), pp. 87–107. ISSN: 1572-8641. DOI: 10.1007/s11023-018-9484-3.

[455] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. ICML 2017 (Sydney, New South Wales, Australia). Ed. by Tony Jebara, Doina Precup, and Yee W. Teh. Proceedings of Machine Learning Research 70. Proceedings of Machine Learning Research Press, 2017, pp. 3319–3328. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.

[456] William R. Swartout. "XPLAIN: A System for Creating and Explaining Expert Consulting Programs". In: *Artificial Intelligence* 21.3 (1983), pp. 285–325. ISSN: 0004-3702. DOI: 10.1016/S0004-3702(83)80014-9.

[457] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. "MoviExplain: A Recommender System with Explanations". In: *Proceedings of the 3rd ACM Conference on Recommender Systems*. RecSys 2009 (New York City, New York, USA). Ed. by Lawrence D. Bergman, Alexander Tuzhilin, Robin D. Burke, Alexander Felfernig, and Lars Schmidt-Thieme. New York, Ny, USA: Association for Computing Machinery, 2009, pp. 317–320. ISBN: 978-1-60558-435-5. DOI: 10.1145/1639714.1639777.

[458] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper With Convolutions". In: *Proceedings of the 28th IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR 2015 (Boston, Massachusetts, USA). Ed. by Horst Bischof, David Forsyth, Cordelia Schmid, and Stan Sclaroff. Piscataway, NJ, USA: IEEE, 2015, pp. 1–9. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298594.

[459] Shinichiro Tabayashi. *An Implementation of Score-CAM With Keras*. The code is available in a Jupyter Notebook. 2021. URL: https://github.com/tabayashi0117/Score-CAM (visited on 10/31/2022).

[460] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY, USA: Knopf, 2017. ISBN: 978-1-101-94659-6.

[461] Richmond H. Thomason. "Indeterminist Time and Truth-Value Gaps". In: *Theoria* 36.3 (1970), pp. 264–281. ISSN: 1755-2567. DOI: 10.1111/j.1755-2567.1970.tb00427.x.

[462] Judith J. Thomson. "Killing, Letting Die, and the Trolley Problem". In: *The Monist* 59.2 (1976), pp. 204–217. ISSN: 2153-3601. DOI: 10.5840/monist197659224.

[463] Nava Tintarev. "Explanations of Recommendations". In: *Proceedings of the 1st ACM Conference on Recommender Systems*. RecSys 2007 (Minneapolis, Minnesota, USA). Ed. by Joseph A. Konstan, John Riedl, and Barry Smyth. New York, NY, USA: Association for Computing Machinery, 2007, pp. 203–206. ISBN: 978-1-59593-730-8. DOI: 10.1145/1297231.1297275.

[464] Nava Tintarev and Judith Masthoff. "Effective Explanations of Recommendations: User-Centered Design". In: *Proceedings of the 1st ACM Conference on Recommender Systems*. RecSys 2007 (Minneapolis, Minnesota, USA). Ed. by Joseph A. Konstan, John Riedl, and Barry Smyth. New York, NY, USA: Association for Computing Machinery, 2007, pp. 153–156. ISBN: 978-1-59593-730-8. DOI: 10.1145/1297231.1297259.

[465] Nava Tintarev and Judith Masthoff. "Designing and Evaluating Explanations for Recommender Systems". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. Boston, MA, USA: Springer US, 2011. Chap. 15, pp. 479–510. ISBN: 978-0-387-85820-3. DOI: 10.1007/978-0-387-85820-3_15.

[466] Nava Tintarev and Judith Masthoff. "Evaluating the Effectiveness of Explanations for Recommender Systems. Methodological Issues and Empirical Studies on the Impact of Personalization". In: *User Modeling and User-Adapted Interaction* 22.4-5 (2012), pp. 399–439. ISSN: 1573-1391. DOI: 10.1007/s11257-011-9117-5.

[467] John Ronald Reuel Tolkien. *The Return of the King*. London, England, UK: George Allen and Unwin, 1986.

[468] Ryan Tonkens. "Out of Character: On the Creation of Virtuous Machines". In: *Ethics and Information Technology* 14.2 (2012), pp. 137–149. ISSN: 1572-8439. DOI: 10.1007/s10676-012-9290-1.

[469] Steve Torrance. "A Robust View of Machine Ethics". In: *Proceedings of the AAAI Fall Symposium on Machine Ethics*. AAAI FS 2005 (Arlington, Virginia, USA). Ed. by Michael Anderson, Susan Leigh Anderson, and Chris Armen. AAAI Technical Report FS-05-14. Palo Alto, CA, USA: AAAI Press, 2005, pp. 88–93. ISBN: 978-1-57735-252-5. URL: https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf.

[470] Gregory Trianosky. "What Is Virtue Ethics All About?" In: *American Philosophical Quarterly* 27.4 (1990), pp. 335–344. ISSN: 2152-1123.

[471] J. D. Trout. "The Psychology of Scientific Explanation". In: *Philosophy Compass* 2.3 (2007), pp. 564–591. ISSN: 1747-9991. DOI: 10.1111/j.1747-9991.2007.00081.x.

[472] Donald M. Truxillo, Todd E. Bodner, Marilena Bertolino, Talya N. Bauer, and Clayton A. Yonce. "Effects of Explanations on Applicant Reactions: A Meta-Analytic Review". In: *International Journal of Selection and Assessment* 17.4 (2009), pp. 346–361. ISSN: 1468-2389. DOI: 10.1111/j.1468-2389.2009.00478.x.

[473] Chun-Hua Tsai and Peter Brusilovsky. "Explaining Recommendations in an Interactive Hybrid Social Recommender". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 2019 (Marina del Ray, California, USA). Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. New York, NY, USA: Association for Computing Machinery, 2019, pp. 391–396. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302318.

[474] Dieter Vanderelst and Alan Winfield. "The Dark Side of Ethical Robots". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES 2018 (New Orleans, Louisiana, USA). Ed. by Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi. New York, NY, USA: Association for Computing Machinery, 2018, pp. 317–322. ISBN: 978-1-4503-6012-8. DOI: 10.1145/3278721.3278726.

[475] Michael Veale and Reuben Binns. "Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data". In: *Big Data & Society* 4.2 (2017), pp. 1–17. ISSN: 2053-9517. DOI: 10.1177/2053951717743530.

[476] Alfredo Vellido. "The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care". In: *Neural Computing and Applications* 32.24 (2019), pp. 18069–18083. ISSN: 1433-3058. DOI: 10.1007/s00521-019-04051-w.

[477] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. "User Acceptance of Information Technology: Toward a Unified View". In: *Management Information Systems Quarterly* 27.3 (2003), pp. 425–478. ISSN: 2162-9730. DOI: 10.2307/30036540.

[478] Jesse Vig, Shilad Sen, and John Riedl. "Tagsplanations: Explaining Recommendations Using Tags". In: *Proceedings of the 14th International Conference on Intelligent User Interfaces* (Sanibel Island, Florida, USA). Ed. by Cristina Conati, Mathias Bauer, Nuria Oliver, and Daniel S. Weld. New York, NY, USA: Association for Computing Machinery, 2009, pp. 47–56. ISBN: 978-1-60558-168-2. DOI: 10.1145/1502650.1502661.

[479] Luca Vigano and Daniele Magazzeni. "Explainable Security". In: *Proceedings of the 4th IEEE European Symposium on Security and Privacy Workshops*. EuroS & PW 2020 (Genoa, Italy). Ed. by Alessandro Armando, Giancarlo Pellegrino, Paolo Prinetto, Frank Stajano, and Lujo Bauer. Piscataway, NJ, USA: IEEE, 2020, pp. 293–300. ISBN: 978-1-7281-8597-2. DOI: 10.1109/EuroSPW51379.2020.00045.

[480] Giulia Vilone and Luca Longo. "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence". In: *Information Fusion* 76 (2021), pp. 89–106. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.05.009.

[481] Andreas Vogelsang and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists". In: *Proceedings of the 27th IEEE International Requirements Engineering Conference Workshops*. REW 2019 (Jeju Island, South Korea). Ed. by Paola Spoletini and Irit Hadar. Piscataway, NJ, USA: IEEE, 2019, pp. 245–251. ISBN: 978-1-7281-5165-6. DOI: 10.1109/REW.2019.00050.

[482] Eric S. Vorm. "Assessing Demand for Transparency in Intelligent Systems Using Machine Learning". In: *Proceedings of the 8th International Symposium on Innovations in Intelligent Systems and Applications*. INISTA 2018 (Thessaloniki, Greece). Ed. by Tulay Yildirim, Yannis Manolopoulos, and Plamen Angelov. Piscataway, NJ, USA: IEEE, 2018, pp. 1–7. ISBN: 978-1-5386-5150-6. DOI: 10.1109/INISTA.2018.8466328.

[483] Kate Vredenburgh. "The Right to Explanation". In: *Journal of Political Philosophy* 30.2 (2022), pp. 209–229. ISSN: 1467-9760. DOI: 10.1111/jopp.12262.

[484] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford, England, UK: Oxford University Press, 2008. ISBN: 978-0-19-973797-0. DOI: 10.1093/acprof:oso/9780195374049.001.0001.

[485] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. "Designing Theory-Driven User-Centric Explainable AI". In: *Proceedings of the 37th Conference on Human Factors in Computing Systems*. CHI 2019 (Glasgow, Scotland, UK). Ed. by Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–15. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300831.

[486] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks". In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. CVPRW 2020 (Seattle, Washington, USA). Ed. by Terry Boult, Gerard Medioni, and Ramin Zabih. Piscataway, NJ, USA: IEEE, 2020, pp. 111–119. ISBN: 978-1-7281-9360-1. DOI: 10.1109/CVPRW50498.2020.00020.

[487] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. "Explainable Recommendation via Multi-Task Learning in Opinionated Text Data". In: *Proceedings of the 41st International ACM Conference on Research & Development in Information Retrieval*. SIGIR 2018 (Ann Arbor, Michigan, USA). Ed. by Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz. New York, NY, USA: Association for Computing Machinery, 2018, pp. 165–174. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210010.

[488] Xochitl Watts and Freddy Lécué. "Local Score Dependent Model Explanation for Time Dependent Covariates". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 129–135. arXiv: 1908.04839.

[489] Rosina O. Weber, Haolin Hong, and Prateek Goel. "Explaining Citation Recommendations: Abstracts or Full Texts?" In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 136–142.

[490] Adrian Weller. "Transparency: Motivations and Challenges". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Lecture Notes in Computer Science 11700. Cham, Switzerland: Springer International Publishing, 2019. Chap. 2, pp. 23–40. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_2.

[491] James Wexler. *Running TCAV*. 2021. URL: https://github.com/tensorflow/tcav/blob/master/Run_TCAV.ipynb (visited on 10/31/2022).

[492] Handy Wicaksono, Claude Sammut, and Raymond Sheh. "Towards Explainable Tool Creation by a Robot". In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2017 (Melbourne, Victoria, Australia). Ed. by David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone. 2017, pp. 63–67.

[493] Michael R. Wick and William B. Thompson. "Reconstructive Expert System Explanation". In: *Artificial Intelligence* 54.1 (1992), pp. 33–70. ISSN: 0004-3702. DOI: 10.1016/0004-3702(92)90087-e.

[494] Bernard Williams. "Integrity". In: *Utilitarianism: For and Against*. Ed. by John J. C. Smart and Bernard Williams. Cambridge, England, UK: Cambridge University Press, 1973. Chap. 2, pp. 97–98. ISBN: 978-0-511-84085-2. DOI: 10.1017/CBO9780511840852.002.

[495] Bernard Williams. "Persons, Character, and Morality". In: *Moral Luck*. Ed. by Bernard Williams. Cambridge, England, UK: Cambridge University Press, 1981. Chap. 1, pp. 1–19. ISBN: 978-1-1391-6586-0. DOI: 10.1017/CBO9781139165860.002.

[496] Travis J. Wiltshire. "A Prospective Framework for the Design of Ideal Artificial Moral Agents: Insights From the Science of Heroism in Humans". In: *Minds and Machines* 25.1 (2015), pp. 57–71. ISSN: 1572-8641. DOI: 10.1007/s11023-015-9361-2.

[497] Jan de Winter. "Explanations in Software Engineering: The Pragmatic Point of View". In: *Minds and Machines* 20.2 (2010), pp. 277–289. ISSN: 1572-8641. DOI: 10.1007/s11023-010-9190-2.

[498] Susan Wolf. "Moral Saints". In: *The Journal of Philosophy* 79.8 (1982), pp. 419–439. ISSN: 1939-8549. DOI: 10.2307/2026228.

[499] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford, England, UK: Oxford University Press, 2003. ISBN: 978-0-19-515527-3. DOI: 10.1093/0195155270.001.0001.

[500] James Woodward and Lauren Ross. "Scientific Explanation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021. URL: https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/.

[501] L. Richard Ye and Paul E. Johnson. "The Impact of Explanation Facilities on User Acceptance of Expert System Advice". In: *MIS Quarterly* 19.2 (1995), pp. 157–172. ISSN: 0276-7783. DOI: 10.2307/249686.

[502] Markus Zanker. "The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System". In: *Proceedings of the 6th ACM Conference on Recommender Systems*. RecSys 2012 (Dublin, Ireland). Ed. by Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand. New York, NY, USA: Association for Computing Machinery, 2012, pp. 269–272. ISBN: 978-1-4503-1270-7. DOI: 10.1145/2365952.2366011.

[503] Carlos Zednik. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence". In: *Philosophy & Technology* 34.2 (2021), pp. 265–288. ISSN: 2210-5441. DOI: 10.1007/s13347-019-00382-7.

[504] Andreas Zeller. *Why Programs Fail – A Guide to Systematic Debugging*. Second Edition. Cambridge, MA, USA: Academic Press, 2009. ISBN: 978-0-12-374515-6.

[505] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. "Building More Explainable Artificial Intelligence With Argumentation". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI/IAAI/EAAI 2018 (New Orleans, Louisiana, USA). Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. Palo Alto, CA, USA: AAAI Press, 2018, pp. 8044–8046. ISBN: 978-1-57735-800-8. DOI: 10.1609/aaai.v32i1.11353.

[506] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" In: *Philosophy & Technology* 32.4 (2019), pp. 661–683. ISSN: 2210-5441. DOI: 10.1007/s13347-018-0330-6.

[507] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning Deep Features for Discriminative Localization". In: *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR 2016 (Las Vegas, Nevada, USA). Ed. by Lourdes Agapito, Tamara Berg, Jana Kosecka, Lihi Zelnik-Manor, Tinne Tuytelaars, Fei-Fei Li, and Ruzena Bajcsy. Piscataway, NJ, USA: IEEE, 2016, pp. 2921–2929. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.319.

[508] Jianlong Zhou and Fang Chen. "Towards Trustworthy Human-AI Teaming under Uncertainty". In: *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. IJCAI XAI 2019 (Macao, China). Ed. by Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019, pp. 143–147. URL: https://opus.lib.uts.edu.au/handle/10453/136189.

[509] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics". In: *Electronics* 10.5, 593 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10050593.

[510] Jianlong Zhou, Huaiwen Hu, Zhidong Li, Kun Yu, and Fang Chen. "Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking". In: *Proceedings of the 3rd International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. CD-MAKE 2019 (Canterbury, England, UK). Ed. by Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar R. Weippl. Lecture Notes in Computer Science 11713. Cham, Switzerland: Springer International Publishing, 2019, pp. 94–113. ISBN: 978-3-030-29726-8. DOI: 10.1007/978-3-030-29726-8_7.

[511] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation". In: *Proceedingas of the 13th IEEE Conference on Computational Intelligence and Games*. CIG 2018 (Maastricht, The Netherlands). Ed. by Mark Winands, Yngvi Björnsson, Jialin Liu, and Mike Preuss. Piscataway, NJ, USA: IEEE, 2018, pp. 1–8. DOI: 10.1109/CIG.2018.8490433.

[512] Olena Zinovatna and Luiz Marcio Cysneiros. "Reusing Knowledge on Delivering Privacy and Transparency Together". In: *Proceedings of the 5th IEEE International Workshop on Requirements Patterns*. RePa 2015 (Ottawa, Ontario, Canada). Ed. by Julio Cesar do Prado Leite, Sam Supakkul, Liping Zhao, and Lawrence Chung. Piscataway, NJ, USA: IEEE, 2015, pp. 17–24. ISBN: 978-1-5090-0122-4. DOI: 10.1109/RePa.2015.7407733.