ETHICS IN THE AI LIFECYCLE

Theoretical Perspectives, Practical Resources and Recommendations

Tampereen yliopisto
Tampere University

This document has been developed in connection to the project Human-centered Artificial Intelligence Solutions for the Smart City (KITE) funded by the European Regional Development Fund (ERDF), The City of Tampere, Business Finland, the Council of Tampere Region, and Tampere University. The content of this document represents the opinions of the author and not necessarily of the project funders.

The author extends his gratitude to the funders of KITE project, his colleagues Arto Laitinen, Kaisa Väänänen, Thomas Olsson, Saara Ala-Luopa, Maria Hartikainen, Anu Lehtiö, Pertti Huuskonen, Jari Kangas, Tero Avellan, Biju Thankachan, Jouko Makkonen, and Esa Rahtu, and everyone who have been kind enough to provide comments and feedback on different sections of this document.

Author information:

Otto Sahlgren
Doctoral Researcher in Philosophy
Faculty of Social Sciences
Tampere University
Tampere, Finland
otto.sahlgren@tuni.fi

ISBN 978-952-03-2777-4 (online)

CONTENTS

RECOMMENDATIONS

RESOURCES FOR PRO-ETHICAL AI PRACTICES

PERSPECTIVES

TABLES AND FIGURES

INTRODUCTION

An overview of the document and a glossary of key terms and definitions

EXECUTIVE SUMMARY

1 Approaching AI ethics from a lifecycle perspective

The design, development, and use of AI systems and other emerging technologies needs to supported and guided by both shared public values ethics and robust regulation serving the protection of human rights and fundamental freedoms. AI systems and the agents responsible for them should comply with four sets of general principles of ethics:

1) Avoiding unnecessary and disproportional harm to humans, non-human animals, and the environment
2) Promoting the good of humans, the well-being of non-human animals, and the flourishing of the environment
3) Protecting freedom and autonomy, and respecting human dignity
4) Complying with applicable laws and regulations, and promoting procedural, distributive and relational justice.

These principles should be followed throughout the entire lifecycle of any AI system from conceptualization to termination of use by implementing multiple layers of protection including, for example, rights for individuals affected by the use of AI systems, a holistic framework for assessing AI systems' impact, and pro-ethical practices for technology development and use in organizations.

2 Regulating AI to protect human rights and fundamental freedoms

There is a need to regulate AI. Joining existing calls for rights-based regulation, we propose that legal regulation of AI systems—including everything from their design and development processes to principles regulating their use—should be centered on the goal of protecting human rights and fundamental freedoms. Rights are non-negotiable and their protection should not depend on the technology or perceived risk in question at a particular case. The relevant question from the rights-based perspective to regulation recommended in this document is not whether a given AI system poses specific risks, respectively, but what can and should be done to protect individuals' rights and their access to those rights. In this vein, we propose three actions for policymakers: First, existing legislation and its enforcement should be improved in order to address salient problems and risks that frequently arise in connection to the use of AI systems, such as discrimination and extractive data practices. Second, rights and regulations—including, for example, an individual right to access evidence and to contest the output of an AI system—should be introduced to prevent and mitigate novel algorithmic harms. Lastly, the use of AI systems which stand in tension with human rights and fundamental freedoms—including, for example, the use of AI systems for indiscriminate biometric recognition in public spaces—should be banned, at least until the protection of those rights and freedoms can be guaranteed.

3 Integrating AI ethics at an organizational, local, and practical level

Legal norms and regulations are necessary but not sufficient to ensure that AI-based products and services are aligned with public values and that they do not violate human rights and fundamental freedoms. Organizations, development and compliance teams, and individual workers need to also be able to comply with established legal requirements as well as principles of ethics that go beyond the legally required minimum of acceptable conduct. Integrating ethics at a practical level requires that commitments to ethics and accountability are visible in organizations' day-to-day practices, however. In particular, there should be explicit and well-defined roles, responsibilities, practices, structures, and norms of behavior within organizations developing or using AI systems. We recommend the following steps to integrate AI ethics at the organizational level of conduct:

1) Organizations should increase awareness about data and technology ethics within their workforce
2) Organizations should actively build a culture of responsibility and incentivize ethical deliberation and behavior
3) Organizations should implement well-defined roles and practices for pro-ethical AI system development and use
4) Organizations should ensure that workers have the skills and tools necessary to identify and respond to ethical issues that arise during design, development, and use.

Operationalizing AI ethics at a practical level is key. We recommend that the legal requirements, principles of ethics, and localized and contextual values to which organizations should adhere are translated into an actionable ethics framework that can support compliance throughout the entire lifecycle of the AI system. Organizations should define ethical targets which their AI systems should achieve, choose methods and/or indicators for evaluating whether those targets are achieved, and establish a set of system specifications and operational safeguards (or user protocols) to ensure that the targets are met during operational phases. The ethics framework—including the specified targets and evaluation metrics and/or methods—should be used to conduct three kinds of regular evaluations concerning AI systems: overall evaluations, targeted evaluations in light a specific ethical value, and trade-off identification analyses aimed at discovering ethical tensions and value conflicts.

4 Practices and requirements for AI ethics and "algorithmic accountability"

Pro-ethical practices and well-defined requirements for AI systems and their use should be established and implemented to ensure that AI systems are developed and used in ways that adhere to demands of public accountability and which protect, rather than infringe, individuals' rights and freedoms. We discuss and propose a set of applicable practices and requirements:

- Prior to deployment, AI system use-cases should be reviewed against applicable standards of acceptability, necessity, proportionality, and assessed in terms of their expected effectiveness to achieve intended outcomes

- Any considerable negative impact—in particular, on human rights and fundamental freedoms, on affected individuals' rights concerning privacy and data protection, on equality and non-discrimination, and on the environment—traceable to the development or use of an AI system should be identified, mitigated and documented prior to system deployment. Impact assessment should also be conducted regularly during the operational phases of AI systems.
- When used in safety critical or other high-stakes decision-making contexts, AI systems should be registered to ensure public transparency and contestability. Other transparency requirements include the requirement for responsible agents to document applied datasets and models, to implement interaction notifications and labels that indicate if and when human users are interacting with an AI system or content generated by one, to integrate capabilities for activity logging and explanation extraction into AI systems, and to engage in truthful and responsible communication concerning AI systems to affiliates, consumers, and the general public.

The previously mentioned requirements and practices would ideally be required by law and guided by a detailed set of standards. However, even if this is not the case either presently or in the future, we recommend that responsible technology developers and operators comply with them for purposes of ethical self-regulation.

WHAT IS THIS DOCUMENT ABOUT?

Artificial Intelligence ("AI")—commonly understood as a field of research but nowadays increasingly also as a family of technologies—is largely considered to be a game-changer for near-future digital societies and economies. Driven by massive amounts of data and computational power, AI systems trained on the past to predict the future can improve our ability to forecast and predict, to optimize processes and operations, to allocate goods and services efficiently, and to personalize what we see, hear, feel, and consume. In many spheres of life, AI systems are also used to structure citizens' behavior and social interactions, to profile people for commercial and safety purposes, to distribute various goods and opportunities, and to aid in different tasks, both mundane and meaningful. Examples of concrete AI applications in these respects include content recommendation and moderation algorithms that operate on online platforms and social media; facial recognition systems used for security purposes and policing; autonomous vehicles that roam the roads, skies, and seas; and social robots that are developed to keep us company, for instance.

Needless to say, the development and use of AI systems presents significant risks. AI, automated decision-making systems, and other emerging technologies risk violating individuals' human rights, restricting their fundamental freedoms, and eroding central values of democratic societies. An abundant research literature and troves of journalistic outputs stand in testimony to the fact that the negative consequences of AI tend to primarily affect marginalized individuals and people who are in existing positions of vulnerability. In the public sector, automated decision-making systems have been used in ways that discriminate against individuals and groups [3], compound and exacerbate existing inequalities [34, 90], and infringe upon individuals' rights [82]. Similar issues can be found in private areas of life, where citizens are subjected to intensifying surveillance [23, 89, 113], and where biased algorithms can affect even people's search for romantic partners on dating apps [54].

The previously mentioned issues only scratch the surface, of course. We might note that the effects emerging technologies can have concern not only individual persons and communities but entire societies and our fragile planet. During the time of climate crisis, in particular, we would be amiss to neglect how the development and use of AI systems creates tangible effects on a planetary scale. AI systems use significant amounts of material resources and energy to enable efficient computation—they are built upon and sustained by massive infrastructures required to transport materials for hardware and system components, to store data, and to supply energy for training and using machine learning models, and to store data, for example [58, 109]. Meanwhile, the "public discourse on AI systematically avoids considering AI's environmental costs" [17]. If the complex environmental impact of developing and using AI-based technologies is left unchecked, novel technologies risk exacerbating the damage we already wreak on our fragile planet on a global scale, potentially preventing us from achieving the most important common end to be pursued in our time—a sustainable future for our planet and life on it.

Unless lawfulness and ethics are understood as key goals of technology development and use, the various benefits AI technologies might offer us are bound to be left unrealized. To prevent violations of legal and moral principles, on the one hand, and opportunity costs due to forgoing what technology would have on offer if done right, on the other hand, lawfulness and ethics must be taken seriously by policymakers, private and public organizations, technology developers, system operators, and civil society. A central challenge is to prevent otherwise beneficial technological innovation and business from being pursued at the cost of human rights and fundamental freedoms. In addition to implementing necessary regulation concerning AI system development and use, there is a also need to integrate ethics deeply into the very culture of the technology industry. Unless these challenges can be overcome, we risk reducing public acceptance and trust and stifling innovation, possibly resulting also in widespread reluctance to adopt novel technologies [25]. Recalling also the planetary scale of effects that technology development and use can have, we should also recognize that the costs of unethical and unsustainable technology can be even more dire, bearing on the prospects of preserving human and non-human life on our fragile planet.

This document has been developed as a part of the project Human-centered Artificial Intelligence Solutions for the Smart City (KITE). The aims of this document are the following:

- to raise awareness about ethics in the context of AI design, development and use
- to discuss pro-ethical approaches to regulating, developing, and using AI systems
- to provide philosophically grounded yet practically feasible recommendations on central issues.

We hope to increase different stakeholders' understanding of ethical issues as they relate to AI technologies and to provide them with ethical perspectives, valuable information, and concrete guidelines on these issues. To encourage sensitivity to the complexity of ethical issues in this context, we want to facilitate "the questioning and reconsideration of any given practice" related to the development and use of AI technology, which requires situating these activities "within a complex web of legal, political and economic institutions" [13]. Given these diverse aims, it might not be entirely inappropriate to consider this document as a "handbook" or a "toolkit" of sorts. We hope it can be used for many purposes ranging from education (both learning and teaching) to supporting the implementation of pro-ethical AI design practices in technology companies. While the target audience of this document includes private businesses and governmental agencies (notably, persons who design, develop, or use AI systems), we hope the document provides useful information for other stakeholders and interest groups as well, such as NGOs, research centers, and citizens. Ideally, the document can help empower citizens of the digital society by helping them critically evaluate applications and technological solutions, to hold both regulators and technology developers to higher ethical standards, and to also criticize and resist technological solutions when necessary. In short, this document is dedicated to all who share our concern for a better future enriched—and not compromised—by emerging technologies.

The contents of this document are organized into various thematically sorted sections. The content of each section is described briefly below. Readers from different backgrounds and with different interests can feel free to pick and choose sections that suit their informational interests.

The first two sections serve as an introduction to AI ethics and our framework. Section 1 provides a narrative overview of recent developments in the field of AI ethics research, discussing three "waves" of AI ethics research within the past decade. The section underlines some characteristic features and approaches of the different waves, including some problematic tendencies identified in connection to each wave. The section may interest researchers and readers interested in AI ethics from a broad, academic perspective. Section 2 presents our proposal for a value framework for AI ethics. We propose a set of fundamental values and four corresponding sets of principles of ethics: nonmaleficence, beneficence, respect for freedom, human autonomy and dignity, and justice and fairness. We define and contextualize these principles to clarify their meaning and philosophical motivations. We also point out some practical implications and discuss examples of risks falling under each principle. The section may interest technology developers and operators, and philosophically-minded readers might be interested especially in the theoretical discussions.

Sections 3-5 discuss different "layers" at which ethics are put into practice. Section 3 starts with the "top layer"—namely, policy and regulation. The section provides some comments and recommendations on regulating AI, and proposes that AI regulation should center on the protection of human rights and fundamental freedoms. We discuss the introduction of novel rights, regulatory mechanisms, and bans on rights-violating systems. The section may be of special interest to readers interested in AI governance and regulation. Section 4 considers the question: how can ethics be integrated into technology development and use at the organizational level? We distinguish four types of pitfalls that organizations should avoid when seeking to integrate AI ethics into their practices. We also provide recommendations on how to raise awareness about ethics in organizations, how to create a culture of responsibility and accountability, how to introduce explicit and well-defined ethics roles and pro-ethical AI practices, and how to ensure that the implementation of those roles and practices results in effective outcomes. The section should be of interest to people working on organizational strategy, operations and management, for instance. Section 5 goes into more detail on the practical aspects of ethical AI and addresses the challenge of operationalizing AI ethics principles in local contexts. We provide theoretical resources and a practical framework for operationalizing ethical values and principles of AI ethics. The framework can be implemented at a practical level to enable responsible agents to monitor morally and legally relevant impacts of AI systems and to empower practitioners to make ethical choices in their work. The section might interest readers working on the ethical aspects of AI systems in industry and research contexts.

Finally, Section 6 proceeds to discuss specific "best practices" for AI ethics and algorithmic accountability that can and should be implemented at various levels. It examines themes such as fairness and transparency and provides recommendations on different types of impact assessment, algorithm auditing, and data and model documentation, for example. We also provide recommendations on how the different mechanisms and practices ought to be implemented and conducted, including by way of providing actionable theoretical and practical resources.

This document offers theoretical and practical resources as well as ethical recommendations. Our recommendations should be taken to consideration while ensuring conformity with constitutional practices, legal norms and regulation, and governing structures of the country or use-context(s) in question, and in conformity with international law, including international human rights law. We hope the document supports compliance and lawfulness in contexts of AI system design, development and use, but we emphasize that the document does not offer a "compliance check" or a "checklist" for compliance requirements. Nonetheless, we wish to also help companies and organizations developing and using AI systems go beyond the "moral minimum" and the requirements expressed by legal norms, by providing insights and describing approaches that can help them seek suitable ways for generating positive impacts on society.

We acknowledge also the political and performative role that ethics guidelines for AI have, and thereby we must clarify that the aim of this document is not to argue for the deregulation of the technology industry. On the contrary, we hold that both (a) improving and enforcing existing legal protections as well as (b) implementing novel regulation concerning AI technologies is necessary to ensure ethical conduct and the protection of individuals' rights and freedoms. The document is motivated by the existence of a genuine gap pertaining to ethics in the technology industry, however—a gap that arises both from meta-regulatory approaches to regulating data and technologies, on the one hand, and a lack of ethics education in engineering and computer sciences, on the other. In this regard, while we do not suppose that self-regulatory or voluntary efforts are sufficient, and that regulation is needed, we wish to provide guidance for responsible agents seeking to implement pro-ethical practices at the ground-level of technology development and use.

A few words also on the limitations of this document. We acknowledge and emphasize that guidelines for ethics in AI cannot capture specific aspects and nuances of a variety of technologies and use-contexts. This is because any set of general guidelines cannot cover all ethically relevant issues and perspectives, and because the development and deployment of particular AI solutions is highly situated and intertwined with local politics, institutional functions and dynamics, and cultural climates, for example. This document focuses on only a limited set of identified and carefully selected ethical questions and areas of concern, respectively. For example, the document presents substantive recommendations on certain key topics—such as safety and security, fairness and transparency—but there are numerous questions that have been left out for pragmatic reasons. We have also primarily focused on providing information and resources that can help different agents develop the necessary capacities for designing and using AI systems in an ethical and responsible manner. These types of higher-order skills and capabilities for ethical deliberation can be highly important—perhaps more so than knowing everything about fairness or transparency in the context of AI, for instance. The document discusses how responsible can establish skills and processes for anticipating and addressing ethical issues, how they should address value conflicts, and what kind of processes and tools they ought to implement to ensure that key values are not compromised, for example. For these reasons, we encourage—and hope to help—responsible agents to actively identify ethical issues that might arise in relation to their specific AI products and services, to treat those issues with due regard for their contextual nature, and to recognize those issues as subject to contesting views and public opinion.

KEY TERMS AND DEFINITIONS

A brief glossary of key terms and concepts is provided next to assist the reader in navigating this document. We also encourage the reader to use existing glossaries of related terms. Let us first address the looming challenge of defining "artificial intelligence" or "AI". The term has traditionally referred to a field of research studying the relationship between information systems and human intelligence—in particular, how to make computer systems mimic or conduct cognitive tasks traditionally taken to require human intelligence. But we will have to bend the term slightly here given that the term "AI" has shifted in meaning during recent decades to cover a set of technological systems or artifacts. We do not have the ambition to pursue a single definition for what "AI" or "AI systems" are. However, here is a tentative definition we will be working with throughout the document:

Artificial intelligence ("AI") covers information-processing technologies that include—and possibly combine—models and algorithms with the capacity to "learn" and to perform different tasks with increasing (yet varying) degrees of functional autonomy. Methods discussed here include, but are not limited to, machine learning methods such as supervised and unsupervised learning, deep learning and reinforcement learning. Computational and cognitive tasks which prove relevant for present discussions include planning, scheduling, knowledge representation and reasoning, search, and optimization, among any other tasks that might involve pattern recognition, correlation calculation, inference, and prediction. Technological systems to which present discussions apply include various types of digital software, where the software can also be embedded in hardware and digital infrastructures, examples including the Internet of Things (IOT), cloud computing services and edge computing, social and service robotics, and human-computer interfaces.

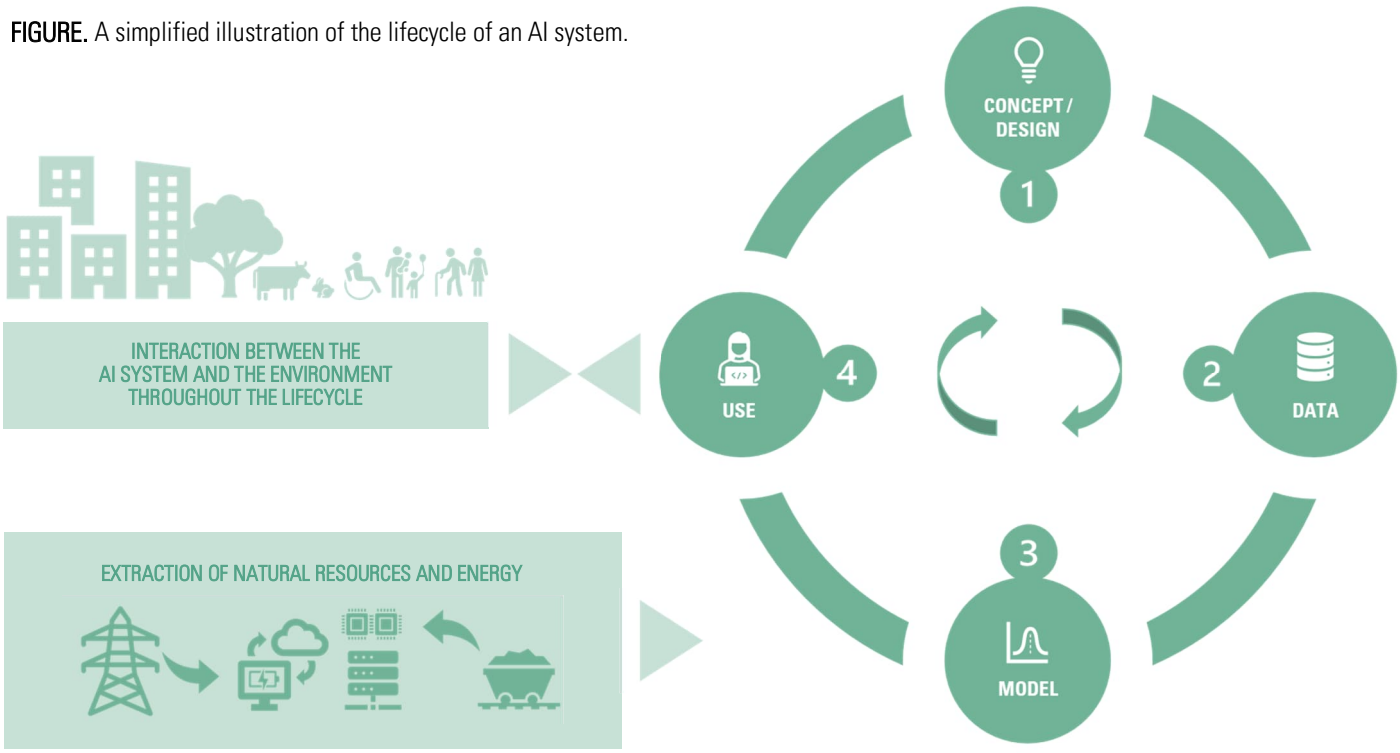
Importantly, the legal and ethical risks that are discussed in this document pertain to many technologies and applications regardless of whether they employ methods that are commonly dubbed "AI". The recommendations that are provided apply in most cases equally to machine learning systems and manually coded software. There might be exceptions because AI systems as they are commonly defined can have specific features or properties which introduce new kinds of problems, or they may exhibit some emergent behaviors which we do not—at least currently—understand. Still, most ethical concerns related to AI tend to only dress age-old ethical problems into new clothing. For the purposes of this document, it is nonetheless useful to focus on both types of ethical concerns: ones specific to (designing and using) AI systems, and the age-old ones.

An algorithm is a series of step-by-step rules that is followed in order to solve a given problem or complete a task. In computer systems, algorithms are the instructions for what the system does computation-wise.

Automated decision-making refers to the use of automatically operating systems, such as computer software, for the purposes of making decisions about natural or legal persons, for instance. Human control and oversight comes in degrees, however. Decisions might be fully delegated to automated systems, for example, or they might merely inform human decision-making to some extent in the specific case. We will use the terms "fully automated" or "automated" and "semi-automated" to indicate relevant differences when and where necessary.

A lifecycle refers to the entire lifespan of a product, service, policy or system, such as an AI system or machine learning model. The term 'lifecycle' is meant to capture not only the central stages included into the traditional software or machine learning model pipeline but also various processes that both precede the development phase as well as the various use-phases of the system, including the termination of its use. The lifecycle on AI system starts roughly from the ideation of a concept or design and the formulation of a computational or data science problem. These activities are followed by design, research an development processes, for example, after which the system is tested, refined, and ultimately deployed. In practice, these stages or phases will overlap and there will be movement back and forth, between model building and testing, for example. The concept of the 'lifecycle' also covers possible processes and activities that are adjacent or less strictly related to the core activities of system development and use. In this sense, the extraction of raw materials and natural resources, development and maintenance of necessary infrastructure and logistics, the procurement of hardware and components, and product marketing and communication processes, for example, are covered by the concept of the lifecycle.

FIGURE. A simplified illustration of the lifecycle of an AI system.



Data means any kind of information captured and stored in representational form, such as digital form. Data can consist of images, video, audio, text, logs and records, statistics, etc. Data can be categorized according to types, such as tabular data and time-series data. Biometric data, which is of special interest in the context of this document, concerns or represents the physical, physiological, or behavioral characteristics of a natural person, and such data can be used (more or less successfully) to detect, recognize or identify natural persons, for example, or to assign them into categories which can be inferred from such data. In the context of AI, one often needs to also distinguish between training data (a set of data used to train the AI system) and evaluation data (a set of data used to evaluate model performance) as well as between input data (data used to prompt or query the system) and output data (data generated by the system as a result of information processing).

A decision-making context refers to the context where an AI system is used to assist or execute decisions, tasks or activities. In this document, we distinguish two types of contexts: First, high-stakes decision-making contexts cover roughly any decision-making or other operational context (either public or private) where use of an AI system can or is likely to have significant impacts regarding whether and how human subjects can access or exercise their fundamental rights and freedoms, or impacts on whether and how they can understand and conduct their lives within a given social context or society. The definition also covers AI systems that comprise safety components of technology products or applications in the relevant contexts. High-stakes contexts include, for example, critical infrastructures, educational or vocational training, employment, workers management and access to self-employment, essential private and public services, law enforcement and administration of justice democratic processes, migration, asylum, and border control management. Second, low-stakes contexts include all contexts other than high-stakes contexts. While this is arguably somewhat vague, low-stakes contexts might include irregular and small-scale entertainment use-contexts, for example, and other contexts where the use of a system does not have significant effects on individuals' access to fundamental rights.

Machine learning is the study of computer algorithms that leverage data and, using pattern recognition and mathematical optimization, "learn" from training examples to improve their performance in a task chosen and defined by a human. Machine learning algorithms are used to train a model based on a sample of "training data". The model is used to make predictions or decisions about novel instances, lessening the need to manually code decision-making rules into the software. There is a variety of different learning techniques that can be applied to train a machine learning model. For example, in supervised learning, the algorithm is shown both training examples and their labels which indicate the correct value of the target variable for a given example (e.g., a correct classification "cat" for an image of a cat). In unsupervised learning, there are no labels and the algorithm is only used to optimize how a set of data is organized into clusters, for example. In reinforcement learning, the learning process is guided by rewarding the system when it behaves desirably and withholding reward (or imposing a cost) when it does not. There are other hybrid approaches as well, which are not mentioned here. Crucially, however, the tasks delegated to the computer system, and the learning process that leads to the model, are determined by humans who formulate the computational problem, select the applied set of training data, and configure the hyperparameters for the learning process.

1 AI ETHICS IN WAVES

A narrative overview of developments in AI ethics

A NARRATIVE OVERVIEW OF RECENT DEVELOPMENTS

In the past decade, the ethical and legal problems related to AI technologies and their use have received increased attention in academic and public discussions. Compared to previous technologies and computer software, AI systems are more adaptive, powerful, faster, complex, dynamic, and can in some restricted cases function without immediate and continuous human control. As a consequence, legal and ethical concerns relating to their opaque nature [4] and other issues including, for example, automated discrimination and increased digital surveillance [23, 34] have become increasingly pertinent. AI ethics as a multidisciplinary area of inquiry has gained more traction as a result, and the research community has gone through significant efforts to make sure smart technologies are built to reflect public values. But what developments have taken place in this field during the recent decade? In this section, we will provide a narrative overview of developments in the field. The overview is not meant to be a systematic presentation of findings, proposed technical methods, or pro-ethical design methods developed by researchers. Rather, by weaving recent developments into narrative form the section provides a brief glimpse into the field and highlights where and why AI ethics—both as a mode of inquiry and a responsible mindset in technology design and development—might go right or wrong. The overview will span three different "waves" of AI ethics summarized below.

The first wave consisted of the development of AI ethics guidelines that introduced more or less coherent lists of ethical values and principles which AI systems should adhere to.

The second wave centered on developing technical methods for addressing ethical issues and challenges identified in the context of AI.

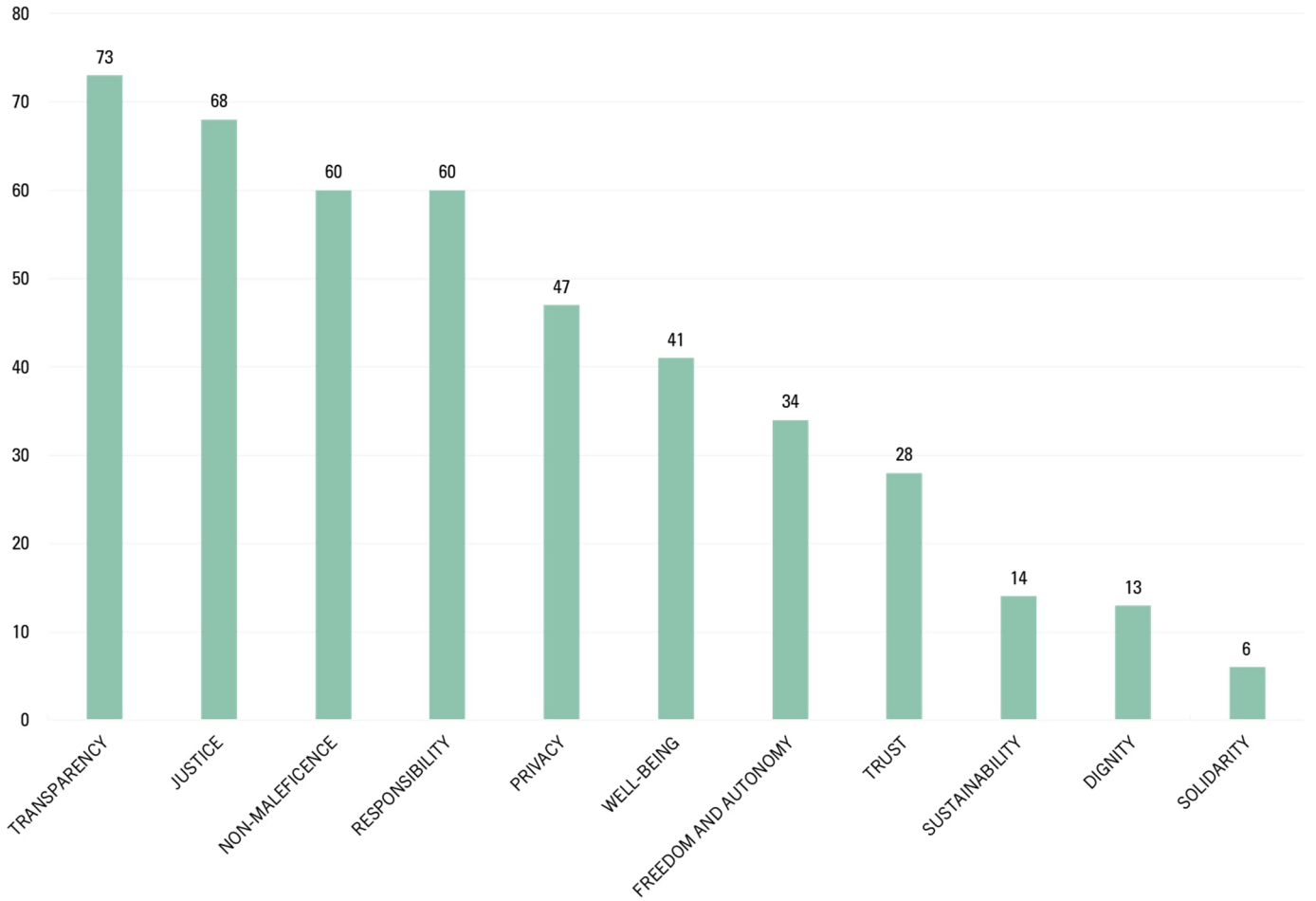
The third wave begun to be challenge the dominant AI ethics discourse, calling to expand the field into more inclusive and multidisciplinary directions.

The idea that AI ethics research has come "in waves" is not new, but the metaphor should not be taken literally. Research projects and orientations are always interrelated and overlapping. In this sense, the waves considered here are perhaps more akin to research streams that continue to flow even today—at times drifting apart, at times joining together. However, we have distinguished them here because their respective characteristics and identified blind-spots can teach us something about the pitfalls we need to avoid and the skills and practices we should cultivate—both individually and collectively as a community.

The first wave

What may be called the first wave of AI ethics consisted of different agents—such as technology companies, research institutes, and non-governmental organizations—developing "ethics guidelines" that proposed ethical values and moral principles for AI systems. The number of ethics guidelines that exists these days is staggering, with repositories (see Linking AI Principles [122] and a repository maintained by AlgorithmWatch) listing over such 160 documents. Reviews of existing guidelines suggest that their content converges on values and principles such as transparency, justice and fairness, responsibility and accountability, benevolence and avoidance of harm, and privacy [57]. The table below, which has been adapted from one review, depicts the content of 84 different guidelines. There is variation in the proposed values across guidelines as well, however, which might be explained by both expectable, general disagreement on ethics and values and differences in the interests that the parties proposing ethics guidelines might have in this context.

TABLE. Ethical principles proposed for AI systems in ethics guidelines N=84)
Table adapted from [57].



Across existing guidelines, there are also many different subprinciples and ethical issues that are mentioned in relation to distinct values and principles [96]. For example, the value of justice is commonly associated with things such as fairness and equality, but also to a variety of things ranging from diversity and value pluralism to accessibility and remedy and redress in decision-making contexts. This suggests that a given ethical principle—justice, fairness or transparency—might in fact mean multiple things within the context of a given set of guidelines, and that the meaning of specific principles might vary across guidelines proposed by different agents.

TABLE. Ethical principles and associated terms in ethics guidelines
Table adapted from [96].

Transparency	transparency, explainability, explicability, und interpretability, comm unication, disclosure, showing
Justice and fairness	justice, fairness, consistency, inclusion, equality, equity, non-bias, non-discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Nonmaleficence	nonmaleficence, security, safety, harm, protection, precaution, prevention, integrity, non-subversion
Responsibility	responsibility, accountability, liability, acting with integrity
Privacy	privacy, personal information, private information
Beneficence	benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	trustworthiness
Sustainability	sustainability, environment (nature), energy, resources
Dignity	dignity
Solidarity	solidarity, social security, cohesion

Guidelines are not enough for ethical AI

The first wave of AI ethics effectively demonstrated a global community's mutual interest in ethically sustainable technology. However, many people in the AI ethics research community were quick to note that ethics guidelines and codes of conduct would not be enough to ensure that advanced technologies would be developed and used responsibly [80, 83]. The proliferation of guidelines and the related focus on values and principles was criticized heavily for various reasons, a few of which are worth mentioning here.

First, most ethics guidelines lack conceptual clarity and philosophical rigor. For example, in most ethics guidelines definitions of core values and principles are vague or lacking entirely. This can lead to false impressions of agreement on values, or a superficial sense of consensus regarding moral principles. Different stakeholders might agree that AI systems ought to be "fair" and contribute to the "common good", for instance, but that agreement is likely to dissolve once they are asked to define what they mean by "fairness" and "the common good". A related issue is that many guidelines confuse intrinsic values (things valued for their own sake such as "well-being" or "justice") with instrumental values (things such as "trust" that derive their value from intrinsic values) or fail to distinguish which values they consider intrinsic or instrumental, respectively [19]. A similar ambiguity also pertains to the use of the concept "responsibility" in many ethics guidelines. Traditionally, responsibility is not considered a value but rather a relation that can have instrumental value. One is responsible for something or someone, or responsible to act in some manner, for example. One can act "responsibly" when they do what they ought to. But it is not valuable to be responsible for wrong-doing in the sense of being the cause of it—one should avoid it! These kinds of conceptual issues are problematic as such, but they can also lead to practical problems since vaguely defined values and principles are difficult to operationalize in practice. Indeed, studies have underlined that designers and developers need concrete tools and procedures, as well as clear standards and requirements, for pro-ethical AI practices [51].

Secondly, many have argued that "tech ethics"—viewed as codes of conduct and other forms of self-regulation—is incapable of ensuring the reproduction of fundamental values where needed, and insufficient for preventing or ceasing salient algorithmic harms when necessary [46, 47, 80, 83]. A common issue that has been raised regarding ethical codes of conduct is that they seem to have little or no effect on practitioners' behavior, for example, suggesting that codes of ethics can be largely ineffective to safeguard, promote, or even assist ethical behavior [47]. This implies that values and principles—as high-level commitments—are not enough to ensure that technology development and use actually adheres to public values and accepted ethical and legal norms. The common sentiment among AI ethicists was—and still is—that the technology industry is in need of enforceable regulation to prevent unethical conduct. While many of the criticisms in this vein tend to mischaracterize "ethics"—viewing ethics as simply a mode self-regulation that companies may voluntarily engage in if they so wish—they do underline a crucial issue: If codes of ethics remain unenforced, violations of those codes are likely to be ignored in practice. Indeed, as noted on the AI Myths website, a majority "of the biggest tech companies are signatories to numerous sets of ethical guidelines, and yet routinely roll out AI products that cause harm".

The second wave

The following wave of AI ethics research centered on developing technical methods for addressing ethical issues, including by way of aligning AI systems with human values as well as moral and legal norms. In other words, during this wave, the research community as well as industry agents focused on developing technical tools, best practices, and evaluation methods, which could be used to "bring ethics into practice". Prominent examples here include tools and metrics that have been developed for the purpose of building fair and explainable AI systems, and which can be implemented across a variety of machine learning methods and AI applications. For example, within a time-span of less than a decade, the field came up with over 20 metrics for fairness with which developers and model builders can estimate different types of biases that AI systems may have learned from their training data [111]. Similarly, a number of explanation extraction tools were developed to provide users information about how a machine learning model generates an output, for example, and about the features that have the greatest effect on the output [81].

In many ways, characteristic to this second wave of AI ethics—a wave still ridden by many with notable success—is the growing recognition of values beyond those traditionally prioritized in technology development, engineering, and business. In other words, it has become increasingly clear that AI systems need to be infused and aligned with values beyond mere "effectiveness", "high performance", and "utility". The primary method envisioned as a means to achieve these goals, notably, was to translate commonly accepted values into metrics, computational methods, and tools which can be applied throughout the AI system lifecycle. The second wave thus answered the needs of practitioners looking for applicable tools for ethical AI, in particular. But problems emerged as well.

Ethics and the lure of solutionism

It quickly became clear that perhaps the most pressing problem with the abovementioned approach lies in the lure of thinking that social and political problems can be reduced into technological ones. Critics of the second wave noted that, in both research and industry contexts, technological "fixes" are often applied without careful consideration of their underlying moral and epistemic assumptions, and the social ramifications that may follow from the application of technical interventions, such as bias mitigation methods. In many cases, the ethical and political problems that specific AI systems bring to the surface (or even exacerbate) are also far too complex and multi-faceted to be solved merely by tinkering with algorithms. The question arose regarding whether ethics discourse in AI is being subsumed by "solutionism" or "technochauvinism": a mindset and ideology wherein the complexity of social and political problems remains unrecognized, and where envisioned "solutions" to social problems are primarily technological in nature [18, 98]. A concern raised by many researchers in this regard pertains to the solutionist fixation on merely removing or repairing "broken parts" of sociotechnical systems which can be fundamentally broken themselves [67]. For example, applying fixes to "de-bias" datasets or algorithms can in worst cases be catastrophic since the problematic tendencies or ideologies underlying the decision-making practices (or institutions) served by the technological system will remain in place even if the technical fixes are successful in some sense.

The third wave

The lesson learned from the second wave of AI ethics was this: technological fixes can in the worst case only make more efficient that which should be reimagined or abolished entirely. An ethical approach to technology, this suggests, requires evaluating any technology under question against the broader backdrop of its social context and, for instance, the institutional or political functions it serves. From a theoretical perspective, AI ethics thereby requires interdisciplinarity and the application of many kinds of moral lenses and theories. The third wave of AI ethics—which might be understood as an "expansionist" movement—sought to expand the range of discourses in AI ethics and to guide both theory and practice into more inclusive and multidisciplinary directions. AI ethics research throughout the third wave has drawn on feminist epistemology and ethics, as well as non-Western philosophies, in developing approaches to ethical technology design and regulation. It has also increasingly sought to question "ethics talk" and shift the focus on the operative forces and interests behind it [46]. Power becomes the key currency here: AI ethics becomes recognized also as a matter of power [59]. It is recognized as the social and political power to "define" the operative parameters and content of "ethics", in particular—a power that is not equally distributed in society. From political organization and policymakers to "Big Tech" companies operating with more or less regard for public accountability, it seems, are currently able to define and enforce which conceptions of "justice" and "social good" become standard by building them into technological systems as well as by lobbying and affecting political processes. The power to define norms and socially enforced standards tends to track various forms of social, political, and economic capital as well as historical divides that have resulted from colonialism and other forms of oppression. To address these issues, and to redistribute the power to define "AI ethics", the field has had to shift its gaze beyond algorithms and data—it has increasingly had also to grapple with complex political realities, past and present.

Each of the previously described waves of research are still going strong. But what can we learn from them? To start, the first wave of AI ethics showed that there is a global interest in ethical AI, although it is apparent that there is also disagreement on what is "ethical". Different stakeholders, such as civil society and the technology industry, have specific interests that should be taken into account when those disagreements are resolved. Once consensus is achieved, the various values and interests need to also be implemented in practice. The second wave has demonstrated that tools and methods for ethical AI can be of assistance here: they allow us to probe and improve AI systems' behavior against various ethical standards. There are no simple fixes for complex social and political problems which AI systems are sometimes hoped to solve, however, as the third wave has shown. We would do well to recognize the limits of technological fixes. What is needed most often is a clear understanding of the addressed problem in all of its complexity, a pinch of humility, and an overall orientation of care towards the individuals who are affected by a given problem or the technology envisioned to solve it. Making room for democratic, deliberative, and inclusive discussion about what technology ideally ought to be like is essential, but there are structural obstacles standing in way of true participation in these discussions—unequal distributions of power, in particular, which should be actively dismantled when and where they occur. In these respects, the ethics of AI—including also the ethics of AI ethics itself—cannot be separated from the political dimensions of AI and technology more broadly.

2 PRINCIPLES FOR AI SYSTEM LIFECYCLES

What values should AI systems reflect?

VALUES AND PRINCIPLES FOR AI SYSTEM LIFECYCLES

Values provide reasons for action. This document emphasizes the crucial and powerful role moral values and ideals play in shaping policy and legislation regarding AI systems and their use, as well as in motivating ground-level efforts to build technology that benefits everyone. Drawing on existing guidelines and research on AI ethics, we propose a general framework for ethics in the context of AI system design, development and use. We will here briefly describe three central facets of the framework.

The proposed set of values and principles of ethics

We propose a set of values and corresponding general principles of ethics that responsible agents should adopt and adhere to when designing, building and using AI systems. These values include the well-being of humans, non-human animals and the environment, human autonomy, freedom and dignity, justice and fairness. We have translated these values into high-level principles of ethics that call for their protection and active promotion by responsible agents: (1) nonmaleficence, (2) beneficence, (3) freedom, autonomy and human dignity, and (4) justice and fairness. To clarify the implications of each principle, we have distinguished three subprinciples under each high-level principle which we also discuss in detail below, highlighting their philosophical motivations and theoretical background. Theoretical resources and practical recommendations regarding the operationalization of the principles is provided also in other sections.

Multi-layer, multi-stakeholder and multi-mechanism protections

A comprehensive and wide-ranging approach which includes appropriate forms of redundancy is needed to ensure compliance with legal and ethical norms and to effectively safeguard the abovementioned values. An ethical approach to AI should be implemented and pursued in a multi-layered, multi-stakeholder and multi-mechanism manner. Protecting fundamental values will require different types of mechanisms and safeguards to be implemented at different levels, and that the relevant responsibilities regarding the implementation of ethics and protection of rights and freedoms should be distributed fairly between stakeholders. The table on the following page illustrates this notion. For example, laws and regulations are required at the level of policy; standards and oversight mechanisms at the level of sectors and industries; roles, norms and practices at the level of organizations; and methods and tools at the level of the aforementioned practices. Any established protections and safeguards should take into account that AI system development and use typically involves a variety of affected stakeholders—for example, regulators, developers, operators and affected individuals—whose interests should be taken into due consideration.

TABLE. AI ethics from a multi-layer, multi-stakeholder, and multi-mechanism perspective

INTER-)NATIONAL GOVERNMENT

- Laws and regulations established by representatives of the public chosen through democratic processes
- Guidelines and standards provided by regulators

INDUSTRIES & SECTORS

- Guidelines and standards specifying industry- or sector-specific requirements
- Codes of ethics and "best practices" for compliance developed with input from domain-experts

ORGANIZATIONS & BUSINESSES

- Ethical and social awareness and a cultivated culture of responsibility and care
- Capabilities, roles, and practices for pro-ethical AI design, development, and use

DEVELOPERS & OPERATORS

- Well-defined and localized objectives and constraints for specific tasks and workflows
- Skills, tools, and learning opportunities for practicing pro-ethical AI design, development, and use

- Enforcing laws and regulations
- Establishing regulations and novel rights for individuals
- Improving individuals' access to human rights and fundamental freedoms
- Oversight and enforcement by regulators and/or authorized bodies
- Supporting independent, high-quality research and journalism around AI and emerging technologies
- Cultivating cultures of ethical innovation and socially responsible business
- Community-building and co-designing with affected stakeholders
- Supporting and engaging in collaboration across scientific disciplines and organizations
- Implementing mechanisms for public transparency and accountability in organizations
- Integrating explicit and well-defined roles and practices around data and AI ethics in organizations
- Educating citizens and providing access to learning resources and opportunities
- Resisting the development and deployment of harmful technologies through citizen activism

A lifecycle approach to AI ethics

A last important aspect of our proposal pertains to the notion that fundamental values—and by extension human rights and fundamental freedoms—should be protected and promoted throughout the entire lifecycle of the AI system, also with due regard for what we call the material and human dimensions of AI. This so-called "lifecycle ethics approach" promotes an understanding of "AI ethics" as an on-going and dialogical, normative activity that begins prior to development of an AI system and extends beyond launching that system. In other words, commonly accepted values and ethical principles should motivate, guide and constrain responsible agents' actions at every step—ranging from the process of conceptualizing an AI-powered product or service to the very end stages of terminating the use of the AI system in question. The approach described here seeks to address two problematic tendencies that characterize dominant approaches to AI ethics. Let us consider them next.

On the one hand, a central problem with traditional approach to ethics in the context of AI is that ethical issues are far too often understood only as pertaining to the software pipeline or to specific stages of that pipeline (such as data collection or model training). There are various ethical questions and problems beyond these activities and tasks that may remain unaddressed if the frame of observation and ethical consideration is not extended to include further stages of the lifecycle, however. Recalling our discussions on "technological fixes" in the previous section, the very activity of framing a problem should be understood as value-laden since some ways of framing a problem tend to prioritize technological interventions as solutions to the perceived problem by default, as it were. On the other end of the lifecycle, we might also note that ethical questions also arise in connection to ceasing the use of a given system. Over time, people may have become dependent on using the system to access certain goods or services (think of assistive technologies for people with disabilities) or to be able to conduct important work tasks (think of software applications that generate images, texts, or designs). The lifecycle perspective emphasizes that these stages, which are not considered in most of AI ethics research or in industry practices, should be also understood as morally relevant.

On the other hand, the lifecycle perspective aims to also overcome the narrow idea that we can build "ethics" and "values" into technological artefacts merely as static rules or behaviors, for example. The "machine ethics" approach is a good example here since its methodological focus is on developing AI systems (or robots) that can follow explicit moral rules and constraints, and to even build artificial moral subjects capable of genuine ethical deliberation and action (see [84]). The methodological benefits of the machine ethics approach notwithstanding, the central premise of this approach can be highly problematic if taken at face value. Specifically, the notion that ethics and values are things that can be simply built into AI systems—for example, through machine learning when we only select the right sets of data and determine the right reward functions for the algorithm—is one step short of attributing (moral) agency to technologies (as opposed to the humans behind them). This is because the notion implicitly frames questions of ethics as if they are simply questions of optimization and rule-following. In other words, the view equates ethical deliberation to "algorithmic thinking" which, many would argue, runs counter to the fundamental idea underlying ethics—namely, that one's deliberation must remain open to diverse perspectives and viewpoints.

The lifecycle perspective to AI ethics aims to overcome the previously discussed limitations of the standard approach. It acknowledges that the lives and places touched by AI systems are not limited to those immediately affected by the data collection processes or the use of the system, but include those who figure into various relations with the systems in a much broader sense. In other words, the approach described here encourages responsible agents to look beyond data and algorithms when engaging in ethical deliberation and problem-solving, and to direct their attention also to the humans, animals and environments that are directly and indirectly affected by the development of an AI system. They should seek to ensure that data workers' rights are respected in those contexts where data is collected for purposes of building an AI system. They ought to also consider whether the "downstream use" of their AI-powered products or services by a third party can be made safer and more transparent. Identifying and responding to ethical issues in the context of AI, in other words, is an activity that should be carried out throughout the entire value chain of the system.

The lifecycle ethics approach also emphasizes the need to consider different temporal perspectives when engaging in ethical troubleshooting and deliberation. This means that responsible agents are encouraged to consider both the history and future of the developed or used AI system—even of specific components, such as training datasets collected or applied for the purposes of model training. Responsible agents might ask who has gathered their training data, with what methods, and in which historical, cultural or political context. They might ask whether data subjects' rights are respected in that context by the party collecting the data, and they should refrain from using data that is gathered unlawfully or unethically. This is important since even some of the most impressive technological innovations in the field of AI tend to involve extractive practices such as outsourced low-paid data labor [44]. Equally important questions might pertain to the future of AI systems and specific components. Responsible agents might consider whether they should allow a third party to access their data, or whether they should be allowed to use the AI system for novel purposes. Also a question concerning the future, an organization deploying an algorithm for decision-making purposes might wish to consider how the algorithm will perform in case the targeted population's base rates suddenly shift, and whether its long-term use might create unwanted effects, such as negative feedback-loops [33].

A last lesson that motivates the lifecycle ethics approach is that software pipelines are messy: they involve interdependent stages and activities, such as data collection and model training, but also interdependent choices. From an ethical perspective to decision-making in the software pipeline, one should note that the choices that are made during earlier stages of the pipeline can affect, enable, and constrain choices that are (or which have to be) made down the line. This is to say that there can be so-called "path-dependencies" between choices—a given decision can either close or open up some possible path that was previously available (or unavailable)—as well as so-called "windows of opportunity"—a given choice to do good or to avoid harm, for example, might be available only for a limited time. Choices made "upstream" affect what responsible agents can and should do "downstream". Hence making the right choices already in the process of conceptualizing the system is crucial, and so is anticipating the opportunity costs those choices might involve at later stages of the lifecycle. The lifecycle perspective is meant to bring these potentialities into responsible developers and operators' field of vision, and to highlight pitfalls on the journey on which they embark when seeking to design, build, and use AI systems ethically.

FIGURE. The lifecycle and material dimensions of an AI system

IMPACTS OF USE & INTERACTIONS

Direct and indirect effects of system use

Short– and long-term Impact of system use

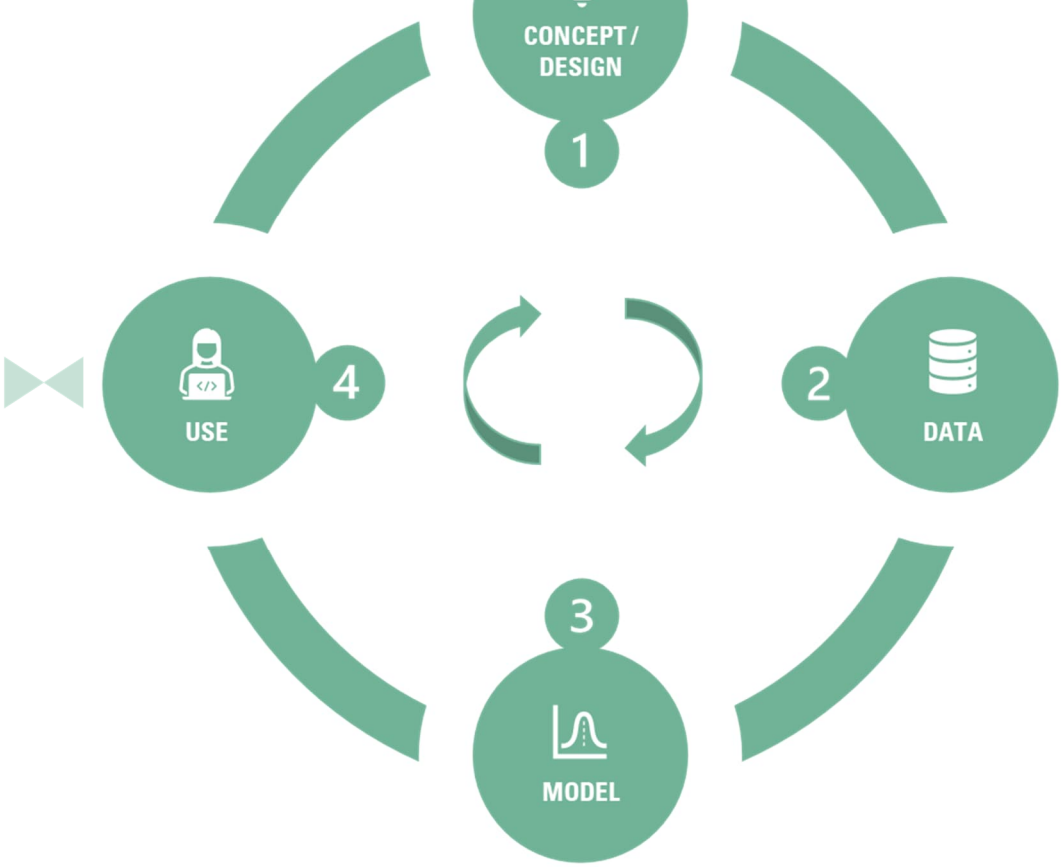
Small– and large-scale effects of deployment

Minor and major impact

One-time or continuous interaction with the AI

Adaptation to system use

Effects of ceasing use of the AI system



EXTRACTION OF NATURAL RESOURCES AND ENERGY

- Natural resources used for material components
- Energy used for model training and use
- Infrastructure such as data centers

1	CONCEPTUALIZATION AND DESIGN	Conceptualization and design of the AI product or service (component) Problem formulation and stakeholder identification Specification of system and operational requirements ...
2	DATA COLLECTION AND PRE-PROCESSING	Data capture and collection Data pre-processing and labeling Data storage and transfer ...
3	MODEL BUILDING AND EVALUATION	Model training and hyperparameter tuning Model evaluation and validation System testing and prototyping (in vitro) ...
4	USE-PHASES OF THE SYSTEM	System testing and prototyping (in vivo) System deployment and use System maintenance and termination of use ...

Before we move on to consider the proposed values and principles in detail, we address some looming concerns. The reader might wonder why the values included in our proposal do not include other values that are (rightfully) mentioned in existing AI ethics guidelines and research, such as privacy and transparency. There are two reasons for this. First, from a theoretical perspective, we consider many values that are mentioned in existing guidelines to be instrumental to realizing what we call "core values" or "fundamental values". As noted in Section 1, fundamental values should be distinguished from instrumental values. Privacy and transparency are highly significant in the context of AI, of course, but they are primarily instrumentally valuable in virtue of the role they play in promoting human autonomy and avoiding harm [66]. That something has instrumental value does not mean that it has less value, nonetheless. Second, our proposal is meant to capture only the core set of fundamental values that should be universally protected—i.e., protected and promoted regardless of the context of AI design and use—and thereby it establishes only a set of minimal requirements, as it were. We have aimed to be concise and clear in defining our framework, but naturally a broader set of values is essential to well-functioning, pluralistic and peaceful democratic societies. Things such as political participation, inclusion, diversity and solidarity—which are not explicitly in our framework yet present in spirit—are good examples. Within limits of reasonable pluralism, in other words, agents can uphold other values provided that they can justify the pursuit of those values to affected individuals.

Another point we emphasize here is that responsible agents who wish to implement our framework should also ensure that they comply with national and international law (including the United Nations Charter) and respect internationally agreed upon objectives (such as socio-political, environmental, educational, scientific and economic sustainability objectives and the United Nations Sustainable Development Goals). This means that responsible agents should operationalize the proposed values and principles in a localized and lawful manner—with due regard for the social context of operation, applicable legal norms and regulations and other relevant normative considerations. In other words, compliance with our proposed framework should not be understood as requiring agents to violate legal norms, regulations, business guidelines, or context-specific ethical codes and safety standards. Rather, we hope that the value framework can complement responsible agents' efforts to ensure compliance with laws and regulations in their specific context. For independent reasons, however, we might also note that the aforementioned norms, regulations, guidelines, codes, and standards should not stand in tension with fundamental values such as well-being or justice.

Lastly, the ethical values and moral principles proposed here are inevitably somewhat abstract or vague. As noted above in section 1, this can be problematic because vague principles can fail to guide action. While a lack of actionability is a genuine concern, we underline the fact that almost all principles—whether moral, legal or otherwise—require practical reasoning and contextual judgment when they are applied in practice. This is to say that no set of principles or guidelines can provide answers to practical and contextual issues that may arise, but they can nonetheless provide actionable guidance by directing agents' attention to specific questions and relevant factors, for instance. To mitigate some expectable problems in terms of action-guidance, however, we will next proceed to discuss the theoretical underpinnings and philosophical context of the proposed principles. We also offer practical recommendations on how to operationalize the principles in other sections below.

RECOMMENDATIONS

Our proposal for principles of ethics for AI system lifecycles

1 NONMALEFICENCE

AI systems should designed, developed and used with both the intention and outcome of protecting humans, non-human animals, and the environment from unnecessary and disproportionate harm.

- ☑ Avoid harm to humans.
- ☑ Avoid harm to non-human animals.
- ☑ Avoid harm to the environment.

2 BENEFICENCE

AI systems should designed, developed and used with both the intention and outcome of promoting the well-being and flourishing of humans, non-human animals, and the environment.

- ☑ Promote the good of human individuals and collectives.
- ☑ Promote the well-being of non-human animals.
- ☑ Promote the flourishing of the environment.

3 FREEDOM, AUTONOMY, AND DIGNITY

AI systems should designed, developed and used with both the intention and outcome of respecting individuals' right to freedom and treating them as autonomous individuals deserving of dignity.

- ☑ Protect and promote affected individuals' freedom.
- ☑ Protect and promote affected individuals' physical, cognitive, and relational autonomy.
- ☑ Respect human dignity.

4 JUSTICE AND FAIRNESS

AI systems should designed, developed and used with both the intention and outcome of complying with any applicable laws and regulations, showing all stakeholders equal concern, treating individuals fairly without discrimination, and promoting relations of equality and respect within society.

- ☑ Comply with all applicable laws, regulations, and standards.
- ☑ Show equal concern for affected stakeholders' needs and interests.
- ☑ Promote procedural, distributive, and relational justice.

NONMALEFICENCE

The proposed framework includes two fundamental principles which appear quite similar: On the one hand, the principle of nonmaleficence states that AI systems should designed, developed and used with both the intention and outcome of protecting humans, non-human animals, and the environment from unnecessary and disproportionate harm. The principle of beneficence, on the other hand, prescribes that AI systems should designed, developed and used with both the intention and outcome of promoting the well-being and flourishing of humans, non-human animals, and the environment. While these two principles are indeed mutually complementary, they are nonetheless conceptually distinct: "Rules of beneficence are typically more demanding than rules of nonmaleficence, and rules of nonmaleficence are negative prohibitions of action that must be followed impartially and that provide moral reasons for legal prohibitions of certain forms of conduct, whereas, by contrast, rules of beneficence state positive requirements of action, need not always be followed impartially, and rarely, if ever, provide moral reasons that support legal punishment when agents fail to abide by the rules" [9]. We will next distinguish three ways the principle of nonmaleficence should be understood in the context of AI system design, development and use.

✔ Avoid physical and psychological harm to humans

The principle of nonmaleficence has its roots in bio– and medical ethics. Etymologically, the principle comes from the Latin maxim "primum non nocere" which means "first, do no harm". The principle prescribes that it is a practitioner's responsibility to act in the best interests of individuals towards whom they have duties of care, to weigh the risks of their own actions against the benefits of those actions, and to take necessary steps to avoid any unnecessary and disproportional harm. Given the origins of the principle in bio– and medical ethics, the relevant types of harm are commonly understood as harms that humans can suffer from. As we explain below in more depth, in our framework this requirement is extended to concern also non-human animals and the environment. But for now, let us discuss the principle of nonmaleficence by considering its specific moral implications more closely in the context of AI system design, development and use. We distinguish three types of requirements that nonmaleficence involves.

☒ Responsible agents should not act with the intention of harming humans with AI systems.

☒ Responsible agents should not unnecessarily or disproportionately harm humans with AI systems.

☒ Responsible agents should not build or deploy unsafe, ineffective or scientifically dubious AI systems.

A first prohibition included under the principle of nonmaleficence is the prohibition on intentional harm-doing. The principle of nonmaleficence prescribes that responsible agents should ensure that the choices made throughout the lifecycle of an AI system—namely, during any phases ranging from system conceptualization to termination of the system's use—are guided by the explicit goal of protecting humans from harm, whether it be physical or psychological harm. In terms of explicit goals, AI systems should be designed to protect patients from diseases, to shield social media users from harmful posts, or to identify and test different types of products for safety risks, for example.

A second requirement is that responsible agents should protect humans from harms that (a) are unnecessary or (b) which do not involve a proportional benefit. This requirement acknowledges that inflicting harm can in certain constrained cases be unavoidable (such as when a medical procedure or a physiological examination involves pain) or even morally appropriate (such as when harm is the appropriate punishment for a crime). The principle nonetheless prescribes that responsible agents should ensure that, even in such cases, the resulting harm in question is not unnecessary, disproportional or otherwise unjustified (when evaluated in relation to the resulting benefit). To adhere to this "clause" of the principle, in other words, one should seek to ensure that any harm (necessarily or justifiably) following from the use of an AI system should be necessary and proportional in relation to the generated benefits.

A third requirement is that responsible agents should not engage in the provision of ineffective or harmful means of treatment. Simply put, responsible agents should not build or deploy AI systems which are ineffective or unsafe. This requirement implies, on the one hand, that responsible developers should ensure that their AI systems meet high standards of performance and safety. From the perspective of an operator of an AI system, on the other hand, it should be made sure that necessary operational and environmental safeguards are in place when the system is deployed. These might include safety protocols that human operators should follow when using the AI system, safety mechanisms such as "kill-switches" and duplicate systems that can be triggered when something goes wrong, and protections implemented in the use-environment of the AI system to prevent and mitigate expectable harms. This last prohibition included under the principle of nonmaleficence also covers so-called "snake oil AI" applications—referring specifically to scientifically suspect and fundamentally dubious applications of machine learning that simply do not and cannot work [85]. Examples of snake oil AI include hiring software systems that allegedly predict a candidate's suitability for a job based on a 30 second video, but the "snake oil" label can be also appropriately attached to most systems that perform tasks such as emotion detection, gender recognition, or prediction of criminal tendencies based on image data. Even if these applications of AI would yield nominally "accurate" results when tested on available datasets, they can be clearly considered scientifically suspect since the relevant indications of performance (e.g., high predictive accuracy on evaluation data) do not track causal connections between the observed data and the target variable—there is no connection between individuals' facial features and criminal dispositions, for example. Fortunately, civil society organizations, researchers, and technology regulators have increasingly started to pay notice to snake oil AI, including AI-powered lie detectors and emotion recognition applications, for example. However, we emphasize that, in the absence of relevant regulations on snake oil, responsible agents have a moral duty to refrain from developing and deploying harmful and/or ineffective systems.

Avoid harm to non-human animals

The discourse around AI and ethics has largely neglected non-human animals and animal ethics perspectives. As researchers Leonie Bossert and Thilo Hagendorff note, "[w]ithin AI research, development and application the role of animals is either not debated at all or it is discussed in the same manner as data usage, thereby implying that animals are nothing more than living resources" [16]. This tendency reflects a broader anthropocentric worldview where the interests, rights, and well-being of humans remain the sole—or at least primary—focus of moral considerations. However, many animal ethicists oppose the exclusion of animals from moral consideration [88, 93] and argue that animals should be recognized, not as mere resources, but as beings with a legitimate interest in not being harmed. Theorists and activists have called for the inclusion of animals into the "moral circle", respectively.

While animals have remained at the margins of discussions on AI ethics, it is clear that the development and use of AI systems can have both direct and indirect effects on non-human animals and their well-being. Research shows that AI technology increasingly contributes to the maintenance and optimization of the so-called "animal-industrial complex" [16]—a term used to refer to the metal cages, fences, breeding and feeding technologies, hormones, and systems of surveillance that are used to restrict and obstruct animal agency with negative effects on their well-being. By optimizing activities such as herd management and fodder delivery for livestock, for example, AI applications are applied to make the animal-industrial complex more efficient—and thereby to also increase the efficiency of various practices that systemically and persistently harm non-human animals. Bossert and Hagendorff summarize the ongoing conflict between increasing automation and animal welfare, noting that "surveillance and AI tools pave the way for not just semi-automated, but fully automated factory farms" which consequentially "leads to even greater emotional distances between the human perception of animal suffering" and incidentally shows all too clearly that the "application of AI on animals is carried out for the benefits of humans, not the former" [16].

Our framework calls for the recognition of the value of non-human animals and their well-being, and calls responsible agents to design, develop and use AI systems in ways that protect animals from harm. The previously discussed three moral requirements—acting without intention to harm, avoiding unnecessary or disproportionate harm, and refraining from building or using unsafe systems—should thereby be extended to cover harms against non-human animals. To be specific, responsible agents—whether it be technology developers or system operators, for example—should not act with the explicit intention of inflicting harming to non-human animals, they should not inflict unnecessary or disproportional harm to non-human animals in case such harms are avoidable, and they should not build or deploy AI systems that create risks for the safety of non-human animals (such as affected wildlife). Taking also into consideration that harm avoidance requires taking active measures, we also emphasize that the development and deployment of AI solutions that explicitly seek to protect animals or entire species from harm should be pursued not only by responsible technology developers and operators, for example, but also encouraged and incentivized at the level of policy.

Avoid harm to the environment

Similar to how AI ethics discourse has largely neglected considerations of animal ethics, environmental sustainability has thus far remained largely in the sidelines in public and academic discussions around AI—some researchers have even pointed out that the "public discourse on AI systematically avoids considering AI's environmental costs" [17]. However, given the severe climate crisis that our planet currently faces, environmental sustainability ought to be adopted as both a key ethical value and constraint in the context of technology design, development, and use. Ensuring that our future of technology development and consumption is compatible with harmonious living with other species on our fragile planet should be an utmost priority. In the present context, considerations of sustainability have two sides [109]: On the one hand, our societies should consider how the use of AI technology could promote environmental sustainability and the well-being of humans and non-human animals on our planet. This notion of "AI for Sustainability" will be considered below in relation to the principle of beneficence. On the other hand, it should be ensured that AI technology is used in a sustainable manner. This is the idea of "Sustainable AI" which we read under the principle of nonmaleficence.

The principle of nonmaleficence includes the prescription that responsible agents should not build or use AI systems that create significantly or disproportionately negative environmental impacts. Rather, responsible agents should seek to identify, mitigate and document the environmental impact of developed or used AI systems. This is a necessary but complex challenge to address, given that the overall environmental impact of a given AI system is constituted by various, distinct sources of impact: On the one hand, there are hardware– and computing-related emissions and impacts [58]. The extent of these impacts is dependent on the amount of data that are used to train the model, the size and architecture of the applied machine learning model, the number of training rounds undergone by the model (also known as "epochs") and the number of times the model is actually used for inference once deployed, for example. The resulting impacts are also dependent on the materials used to build the hardware and the energy infrastructures involved in both building and powering a given AI application, for example, ranging from things such as energy supply and power systems to utilized data centers and personal devices. On the other hand, there are also application-dependent impacts—both direct and indirect—which result from the use of AI applications. On the side of immediate effects, AI technology is already used for various tasks that improve the sustainability of existing operations or mitigate the effects of climate change, for example, but such applications unfortunately constitute a minority of all areas and tasks where AI solutions are currently operating [95]. There are also less tangible but equally important effects that can arise in an indirect manner. For example, the introduction of an AI application can result in a "lock-in" effect, where less carbon– or energy-intensive products are de facto prevented from entering the market due to increased demand of another product—as an example, autonomous vehicles might become the go-to option for transportation instead of a more sustainable option such as mass transit [58].

We will provide practical recommendations on assessing and mitigating the environmental impact of AI systems below. For now, we will move onto consider the principle of beneficence.

Harmful core objective or outcomes

An AI system is intentionally used for inflicting harm or the use of system has the unintended outcome of harming individuals, animals or the environment. For example, a third-party that procures the system uses it for malicious purposes.

Disproportionate or unnecessary harms

An AI system generates harms that are not necessary for achieving an otherwise desirable objective. Alternatively, the benefits that result from the system's use are not proportional to the related costs or generated harms.

Physical or psychological harm

An AI system inflicts unnecessary physical damage or harm to humans or animals, or the behavior or output of the system generates emotional harm or induces anxiety.

Bad performance

An AI system does not perform well or is prone to malfunction. For example, the system makes mistakes when it is presented with novel cases, when it is used in new environments, or when it is used by inexperienced users.

Lack of dependability

An AI system does not meet the standards or requirements of its intended use. For example, it frequently exhibits unexpected or problematic behaviors, or human operators of the system do not know how to constrain the system's behavior.

Lack of resilience

An AI system performs inconsistently at the face of uncertainty, complexity and change. For example, it performs badly when presented with out-of-distribution examples (anomalous data or outliers) or when used in changing environments.

Reproducibility issues

The behavior or outputs of an AI system (such as its predictions) cannot be reproduced. Alternatively, the system fails to meet basic standards of scientific integrity in that it involves false methodological or empirical assumptions, for example.

Model or system vulnerability

An AI system exposes vulnerable individuals to harm due to lack of data security or model security. For example, the system is vulnerable to model attacks used to infer sensitive information or adversarial perturbations used to fool the system.

Lack of human oversight

An AI system is operated without appropriate forms of human oversight. Alternatively, the human operator of the system does not know whether, when or how to intervene on the system's operations.

No "fallback plan"

An AI system lacks a safe procedure for modification, a safely triggerable "kill-switch" or a duplicate system that can be deployed if the primary system malfunctions.

Issues with the system's use-environment

An AI system's use-environment lacks appropriate safeguards and safety control mechanisms. For example, malicious agents' behavior or naturally occurring changes in the system's use-environment create unexpected or uncontrollable safety risks.

BENEFICENCE

The principle of beneficence prescribes that AI systems should designed, developed and used with both the intention and outcome of promoting the good of humans, non-human animals, and the environment. The principle of beneficence is a rather broad principle in that it encompasses "all norms, dispositions, and actions with the goal of benefiting or promoting the good of other persons" [9]. We will look at three distinct sets of moral requirements grounded in this general principle that responsible agents should adhere to in the context of AI system design, development and use. However, before we proceed, we should note that "doing good" can mean various things in ordinary language. For example, often when people talk about "goodness" they might use the term in a pluralist sense of something being good overall or acceptable. However, within the context of our framework, the concept of "doing good" is understood in a more technical manner which also slightly differs from some ordinary interpretations of "doing good".

- Beneficence as promoting the good of individuals and society. In the context of the principle of beneficence, "goodness" is traditionally understood in terms of what is good for humans. These might include things such as well-being, pleasure, happiness, preference-satisfaction, flourishing, and so on. Accordingly, the principle of beneficence states that agents should benefit individuals, communities or society as a whole in these respects. For example, "doing good" in a medical context would be understood as the prescription to do things that increase or protect human well-being, whereas "doing good" might in the context of entertainment be better understood in terms of creating feelings of joy.

- Beneficence as supererogation. Whereas nonmaleficence prohibits the agent from inflicting unnecessary or disproportional harm, the principle of beneficence encourages actions which are morally praiseworthy but not required—"supererogatory" acts that promote the good of other people and the community. Adherence to the principle of beneficence can thus involve going beyond one's duties in a strict sense. Accordingly, the principle "connotes acts or personal qualities of mercy, kindness, generosity, and charity" in addition to things such as "altruism, love, humanity, and promoting the good of others" [9].

Importantly, our framework also extends the scope of relevant beneficiaries to include non-human animals and the environment, meaning policies and technologies, for example, should also benefit animals and our planet.

The previously mentioned interpretations of the notion of "doing good" were put forth to emphasize that responsible agents who are seeking to "do good with AI" should always critically reflect, evaluate and specify what they understand by "doing good". This is highly important as disagreement about what is good is the rule rather than the exception. Yet the application of AI to a given task is often simply assumed to be beneficial as such [45, 96]. In particular, far too often technology developers seem to take for granted that, even if AI technologies cannot provide "perfect solutions to social problems", their use can nonetheless do good "by making many aspects of society better" [45]. This assumption is not true, of course—there are many applications of AI that are harmful (even though they might be developed with good intentions). Highlighting this problematic tendency, researchers have noted "doing good with AI" risks transforming into a slogan used to merely legitimize the development of problematic automated technologies [45]. Bearing this problem in mind, we will proceed to consider what beneficence as a principle of ethical AI could mean in practice.

Promote the good of human individuals and collectives

The principle of beneficence prescribes that AI technology should be developed and used explicitly for the purpose of promoting the good of humans—be it individuals, groups, communities or entire societies. But what exactly does it mean for AI systems to promote "the good of humans"? What constitutes the good of humans is subject to age-old philosophical debates, of course. Drawing on different philosophical frameworks of well-being and happiness, we might distinguish at least four different types of answers [27, 49]: the good of humans could be understood as the satisfaction of desires or preferences, as experiencing pleasure and joy, as well-being or happiness, or as flourishing or living a virtuous life. While the philosophical debate remains, any AI system that is designed and used to promote the good of humans will likely have to contribute to the aforementioned things.

The previously mentioned objectives and goals for AI system design, development and use are largely individualistic in that they focus on the good of individual human beings. Beyond promoting the good of individuals, however, the principle of beneficence underlines the importance of selecting socially and globally beneficial use-cases for AI. Prominent examples of use-cases that would be aligned with the principle of beneficence in this respect include, for example, the Sustainable Development Goals (SDGs) and their corresponding 169 targets proposed by the United Nations as part of the 2030 Agenda for Sustainable Development. As demonstrated both by projects such as AI4Good and recent research [105, 112], AI systems can indeed be applied in ways that can positively impact many SDGs either directly or indirectly. Application of machine learning technology for the purpose of achieving SDGs should be done with care, however, since there are also risks for negative impacts. For example, due to the interdependent nature of SDGs, AI systems that are applied to achieve a given SDG may also have a negative effect on other SDGs [105].

Promote the well-being of non-human animals

Discussing the principle of nonmaleficence, we noted that applications of AI technology can generate and exacerbate harm to animals. Here we note that the principle of beneficence calls responsible agents to design, develop and use AI technology specifically to promote the well-being of animals. An important way to do this is to research and identify ways in which machine learning methods could be applied for the purpose of protecting wildlife from human-generated harms, protecting biodiversity and endangered species from becoming extinct, and so on. The possibilities are vast in these respects, as research shows, as there are both indirect and direct ways to promote animal welfare by applying data science and machine learning. Consider indirect effects first. For example, AI systems have potential to indirectly support animal welfare by decreasing reliance on animal experiments. AI-based testing can be a more reliable way of testing the safety of medical substances when compared to animal experiments [72], for example, and thus can "in the long run […] lead to a drastic reduction of animal experiments due to advances in AI development and use" [16]. In the food industry, AI-based tools have also proven apt for researching and developing plant-based food products, thereby assisting in efforts to transition away from animal products (see [16]). A wide range of existing applications also demonstrate that AI systems can be used specifically and explicitly for the purpose of promoting animal welfare [16]. AI systems can be applied to prevent animals from being harmed by human technologies and infrastructure, for example, as well as to help stray and shelter animals be adopted into suitable homes and prevent them from being abandoned [63]. In an illustrative case of applying data science and AI for the explicit purpose of promoting animal welfare, an AI system called PAWS (Protection Assistant for Wildlife Security) was trained by researchers to predict likely sites of poaching and thereby prevent poachers from harming wildlife [64].

These and other types of applications of AI for purposes of protecting animal welfare should be identified, researched and properly funded. Ideally, non-governmental organizations devoted to protecting animals, for example, should collaborate with technology developers and researchers to identify effective ways to preserve and protect endangered species and wildlife, and to contribute to other worthwhile causes promoting animal welfare.

✔ Promote the flourishing of the environment

Previously it was noted that the development and use of AI systems can have a significant and unfortunately negative impact on the environment, which means that evaluating and improving the "sustainability of AI" should constitute a key ethical priority for responsible agents [109]. However, identifying and mitigating the negative environmental impact of AI systems is a considerable challenge. To succeed in this challenge "the research community needs to develop a holistic and operational understanding of the different ways in which ML [machine learning] can positively and negatively impact climate change mitigation and adaptation strategies" [58]. We will discuss some fruitful approaches in further sections below. In this section we focus on the possibilities that data science and AI can provide in terms of improving the sustainability of existing practices and processes across different application domains and in terms of contributing to various efforts (both local and global) to protect the environment and mitigate the effects of climate change. For example, recent works on AI and sustainability have compiled comprehensive lists of ways how emerging technologies such as AI can help mitigate the negative environmental impact of existing technologies and human activities, and how they can assist in the prevention or mitigation of harms related to extreme environmental phenomena, such as storms [53, 95].

If there ever was a good use-case for AI technology, it is addressing the climate crisis [24, 105]. It is recommended here that AI systems should be explicitly designed and proactively used to address the global climate crisis and its effects, to enhance ecosystem and biosphere sustainability, and to support the transition to clean energy and sustainable industry and lifestyles. In addition to the previously cited evidence concerning the potential of data science and AI in the domain of sustainability, projects such as AI4Good and AI4Climate have also shown promise in relevant areas. Importantly, however, research into how data science and AI solutions could be harnessed to tackle the climate crisis and to mitigate its effects animals ought to be further encouraged. This requires actions on various levels—such as allocating research funding, implementing political instruments and financial incentives, and encouraging a cultural shift towards a greener technology industry. This has the concrete implication that policymakers should allocate and increase funding that goes into research operating at the intersection of computer science and climate science to fuel the development of technological means to address the climate crisis and its side-effects. Research projects and other collaborative efforts to build AI systems that address the existential risks brought about by climate change and environmental degradation should involve interdisciplinary groups of experts, however. Climate and environmental science experts, political scientists, as well as climate and environmental activists should be involved. It should be emphasized that the creation of technological solutions for addressing climate change and its effects should be based on realistic expectations of what AI and other emerging technologies can offer in this respect—it is clear that climate change cannot be prevented with mere technology. In other words, while we would be amiss not to recognize the possibilities of AI in the relevant areas, "AI" should not, at a cultural and discursive level, transform into what philosopher Mark Coeckelbergh calls "an alienation machine"—namely, "an instrument to leave the Earth and deny our vulnerable, bodily, earthly, and dependent existential condition" [23].

"Automation solutionism"

The automated or to-be-automated task necessitates "human touch" and oversight. Alternatively, the goal established for the AI system cannot be achieved with (purely) technological interventions.

Failure to prioritize goals or tasks

There is a more important objective, goal or collective end which should have been prioritized. For example, the AI system is used to achieve a goal that was not as important from the perspective of collective ends or the common good.

Alternative solutions

There is a non-technological and more desirable way to reach the objective or achieve the established goal which does not require a (purely) technological intervention.

Failure to understand the problem

The problem that automation is envisioned to address has not been properly understood in all of its complexity. For example, there are structural, political, economic, historical or social causes that have not been accounted for.

Failure to understand the context

The context in which the AI system is envisioned to operate has not been properly understood in all of its complexity. For example, there are norms, hierarchies, behaviors, values, or other factors which should be taken into account.

Failure to understand stakeholders' interests and needs

The reasonable and legitimate interests of all affected stakeholders have not been taken into consideration. For example, developers have not considered or inquired what affected stakeholder groups consider "good" or "beneficial" for them.

Lack of well-defined beneficial impact

The intended or expected benefits of building or using an AI system are not defined clearly. Alternatively, there is no evidence that building or using the AI system will generate the expected benefits, or the available evidence is insufficient.

Biased problem formulation

The way the problem that the AI system is supposed to solve is formulated manifests considerable bias against a given stakeholder group, a social or demographic group, a given culture or worldview, a political ideology, etc.

"Reward-hacking"

An AI system learns harmful, deceptive or otherwise dangerous behaviors due to a vaguely defined objective function or because the training process lacks constraints that prevent the system from learning "shortcuts" for maximizing its reward.

Disregard for animal welfare

An AI system is used to pursue a goal that is feasible and desirable from the perspective of humans, but the pursuit of the goal generates harms to non-human animals such as endangered species or local wildlife.

Disregard for environmental sustainability

An AI system is used to pursue a goal that is feasible and desirable goal from the perspective of humans, but the pursuit of the goal creates negative environmental impacts by increasing emissions or contributing to environmental degradation.

Disregard for indirect effects

An AI system is used to pursue a feasible and desirable goal without considering the indirect effects and long-term impact that result from the use of the system, such as its effects on community relations or local markets.

FREEDOM, AUTONOMY AND DIGNITY

The principle of freedom, autonomy and dignity prescribes that AI systems should designed, developed and used with both the intention and outcome of respecting individuals' right to freedom and treating them as autonomous individuals deserving of dignity. In other words, the principle underlines that responsible agents—for example, those designing, developing, or using AI systems—owe a duty of respect to each person affected by the system (including those affected by specific processes related to building the system, such as data collection). The minimum requirement of respectful treatment is that the affected individuals are treated in a dignified way without violating their fundamental freedoms or personal autonomy. The three central concepts at play—freedom, autonomy and dignity—in this sense constitute three facets or dimensions of rights that moral persons have and which should not be violated at least without appropriate justification. From an ideal point of view, we would hope nonetheless that responsible agents go beyond the minimum and actively promote the freedom, autonomy and dignity of persons. To provide a better sense of what the principle implies, let us take a closer look at the three key concepts and their implications.

Protect and promote freedom

The first central concept here is freedom. A person or agent is considered free in the general sense when they lack constraints (or preventing conditions) to do or become certain things [20, 73]. Violating a person's freedom consists of unduly limiting it by imposing such constraints without appropriate justification. However, there is debate concerning whether freedom should be understood (a) in a negative sense, where freedom refers to an absence of something (such as an obstacle, barrier, constraint or interference from other persons) or (b) in a positive sense where freedom requires the presence of something that makes one free (such as control over something, having means to an end, or being able to self-realize) [20]. These two conceptions of freedom have different sets of moral implications. Whether freedom is to be properly understood in the negative or positive sense in a given context thereby bears on what is needed from an agent—or an AI system—to respect a person's freedom. From a narrow perspective to negative liberty, for example, a recommender system should not (prima facie) affect what an individual user chooses to do—it should not engage in "nudging" by manipulating the individual's choice environment or the choices they have available. Here, protecting affected individuals' freedom means that the system is prevented from unduly interfering with those individuals' actions. However, from a positive liberty perspective, "nudges" could be also understood as means to promote individuals' liberty in case nudging offers them better means to do what they wish—to consume content that they prefer, for example. Similarly, an AI system which allows users to control the system's behavior of (what it does and what it does not do) can empower individuals from this perspective of positive liberty.

✔ Protect and promote cognitive, physical, and relational autonomy

Freedom is closely related to the second concept of interest: personal autonomy [31]. This type of autonomy is traditionally understood as the human capacity and exercise of self-determination, self-rule or self-governance. While autonomy comes in degrees, an autonomous individual in a general sense has the capacity to rule themselves—to be the author of their own thoughts, values, actions. Personal autonomy in this sense also relates to authenticity in choices and actions—a self-governed individual conducts acts in a way that reflects their own values and commitments. Importantly, one can distinguish cognitive and physical dimensions of autonomy: On the one hand, autonomy concerns an individual's cognitive capacity to form, organize, and evaluate their own thoughts and beliefs, and to make decisions according to their own values and goals. From a physical perspective, on the other hand, an individuals' capacity to maintain their bodily functions and to exercise their mobility through the exercise of their own will is crucial to their autonomy. Persons are always embedded into social contexts and relations, moreover, which means their autonomy is configured by those contexts and relations—an individual's autonomy is developed and exercised in relation to one-self and others, which underlines the relational dimension of autonomy [104]. For example, whether one has a healthy self-image, whether one has supportive social relationships, and whether one receives social recognition from others can affect the development and exercise of their autonomy.

Violations of cognitive autonomy may, as they may occur in the context of AI, include interference with a person's autonomous deliberation through deception or preventing their actions or choices through coercion, for example. The use of an AI system should not lead to deception, manipulation or coercion of individuals, for example. From the perspective of protecting physical autonomy, embedded AI systems—such as social robots and autonomous vehicles—should not obstruct individuals' physical autonomy by interfering with their movements. But given the complexity of human autonomy, there are also numerous other ways in which AI systems can (be used in ways that) violate a person's autonomy [66]. For example, an AI system might not recognize salient aspects of an individual's social identity. Or, conversely, it might treat that individual as a mere member of a group by giving only certain features of the individual weight. In both of these cases, the system fails to express adequate recognition respect for the individual's unique characteristics and individuality (see [28]). From the perspective of informational self-determination, which refers to how individuals exercise their autonomy by accessing information and controlling what pieces of information about themselves others are allowed to access, we might also note that privacy and transparency are considerably important. On the one hand, to exercise their autonomy, individuals need information that is relevant to assessing their own situation. In the context of automated decision-making, for example, this implies that system transparency and the explainability of automated decisions are both central to protecting personal autonomy. On the other hand, to allow individuals to act according to their own values and beliefs without undue pressure, they should be shielded from undue surveillance by others. This implies that questions related to privacy and security are also of key importance. We thereby emphasize that transparency and security—as features or attributes that AI systems may possess to different degrees—derive their value partly from the fact that they are essential to respecting personal autonomy, freedom and dignity (and for protecting individuals from harm).

Respect human dignity

The third central concept here is human dignity [94]. Human dignity is traditionally considered to be universal in that it is ascribed to all humans and inalienable in that simply being a human person grounds the right to dignity. The notion of dignity is, for these reasons, also intrinsically related to the notion of inalienable human rights, such as the right to life and the right to be free from slavery. The moral requirement to respect a person's dignity is grounded in the notion that human individuals are unique and separate moral persons that should be valued for their own sake, and that they have a right to be treated in a dignified manner, accordingly.

What risks might (the use of) AI systems involve in relation to human dignity? We might distinguish at least three types of risks of undignified treatment (see also [66]):

- The (use of the) AI system humiliates or dehumanizes a person, or disparages or denigrates them.
- Persons are treated instrumentally by (the operator of the) AI system—as mere means to an end.
- Persons as treated as interchangeable or in an objectifying manner by (the operator of the) AI system.

Evaluating whether an AI system might pose risks in these manners can be complicated as it requires paying close attention to both the social context and the particularities of different persons: On the one hand, treatment that is disrespectful of a person's dignity can depend on the frame of social meanings within which the treatment takes place. An illustrative example is a case where Google's image processing algorithm labeled an image of a black couple as "gorillas". Whereas labeling a human being as an animal can be dehumanizing as such, this particular case was arguably made worse by the racist social and historical meaning that is attached to acts that relate blackness with the particular label in question. On the other hand, AI systems deal with generalizations and statistical correlations, which implies they will always treat individuals "as mere numbers" or as members of a certain class (such as a gender group or an age group). In this sense, there should be a way to ensure that the particular characteristics of each individual and the social context of system use can be taken into account.

Having discussed these three central concepts, we might note that, while freedom, autonomy and dignity are closely related, they nonetheless ground distinct (albeit overlapping) sets of rights to be respected. This is important to recognize in technological contexts as well, because trade-offs and value tensions can arise in these areas. A given form of treatment that is in principle respectful of an individual's autonomy or freedom, for example, can still fail to express appropriate respect to their dignity as a moral person. To provide an extreme example, an individual might consent to being tortured or publicly humiliated by others, but such treatment would not respect for their value as a human being. In other words, these moral obligations can in principle draw into opposite directions.

Interference with (capacities for) cognitive or physical autonomy and freedom

An AI system obstructs a person's reasoning or deliberation through deception or manipulation, for example. Alternatively, the system interferes with the person's physical integrity or bodily capabilities by preventing their movement, for instance.

Interference with autonomous, free choice

An AI system lacks meaningful consent mechanisms for opting in and out of use. Alternatively, An AI system limits the range of meaningful choices or options a person has available to them (or eliminates them entirely).

Promotion of heteronomy

Addictive behavior or overreliance on the AI system is incentivized, including by virtue of a lack of appropriate safeguards and guidelines for using the AI system in a given context or on a given group, such as children.

Disregard for metapreferences

An AI system caters only to the immediate preferences expressed by persons (such as the desire to purchase new clothes) but disregards their long-term goals and objectives (such as their commitment to more sustainable and less frequent consumption).

Erosion of relational autonomy

An AI system negatively affects a person's social relationships or self-image. For example, the system subjects a person continuously to content which results in the person perceiving themself as of comparatively lower self-worth or as inadequate.

Prevention of informational self-determination

An AI system provides insufficient means for the human to govern their data or their digital representation, or lacks them entirely. For example, it lacks controls that enable the person to decide what personal data is collected, stored and transferred.

Lack of transparency

An AI system does not provide a person sufficient information about the reasons for—or about the factors that affect—a decision which has significant consequences for that person.

Violation of privacy

An AI system excessively and unduly surveils a human person—including, for example, their choices and actions. Alternatively, the system (or its operator) has undue access to sensitive or personal aspects of a person, their identity, choices or actions.

Misrecognition and lack of recognition

An AI system fails to recognize—or explicitly denies—a subjectively or socially salient aspect of a person's identity. Alternatively, the system makes an incorrect inference about a subjectively or socially salient aspect of a person's identity.

Failure to treat as an individual

An AI system treats a human person as merely "a number" or a member of a group. The system does not use sufficiently granular or contextual data about the person, for example, or the output of the system represents a harmful stereotype.

Undignified treatment

The behavior or output of an AI system humiliates, denigrates or derogates a human person. Alternatively, the behavior or output of the system is such that no human should be subjected to it.

JUSTICE AND FAIRNESS

Justice is a central value of political ethics, the realization of which requires that there are institutions that ensure each individual is protected and treated as they deserve to be treated. As philosopher Iason Gabriel has noted, the "key principles that govern the fair organization of our social, political, and economic institutions also apply to AI systems that are embedded in these practices" [41]. That is, if AI systems are to be used in public contexts—for example, to optimize institutional processes or to automate practices that significantly affect citizens' lives and prospects—such systems should arguably be aligned with principles of justice. In our framework, this is taken to mean that AI systems should designed, developed and used with both the intention and outcome of complying with applicable laws and regulations, affording stakeholders due and equal concern, treating individuals fairly without discrimination and with due regard for procedural and distributive demands of justice, and promoting relations of equality and respect between individuals and groups. We will next consider these different constitutive aspects of "algorithmic justice" and their corresponding regulative principles.

Comply with applicable laws, regulations and standards

In most contexts, justice is primarily discussed in terms of lawfulness. The first prescription included under the high-level principle of justice and fairness in our framework prescribes that responsible agents should ensure that AI systems—including also any relevant processes related to their design and operation—are in compliance with laws and regulations established by democratically elected representatives of the public. Lawfulness as a principle of justice for the context of AI means quite simply that AI applications—such as software products or technology components—and the activities in which they are built and used should meet any applicable standards specified in (inter)national laws and regulations. The aforementioned include also safety standards and other requirements that can be industry– or sector-specific, for example, or which apply only locally in some contexts. As the ultimate aim of laws and regulations is the protection of individuals and the public good, we emphasize that lawfulness also refers to the central requirement to respect individuals' fundamental rights and freedoms. Of particular importance in this regard are human rights and the fundamental rights of citizens, which establish essential and concrete protections for citizens. Understanding one's duties and obligations in terms of compliance is integral in this respect: for example, an agent might have not only duties to refrain from violating individuals' rights and freedoms, but also positive duties that require them to actively improve people's access to those rights. If so, AI systems should be used in ways that reflect such positive duties and obligations that agents may have. In other words, and quite concretely, AI systems should in some cases be proactively used to improve individuals' access to their rights and freedoms, for instance.

Show equal concern for stakeholders' needs and interests

A second prescription concerns the demands of justice in the context of technology design and related activities and procedures, such as stakeholder identification, management and engagement, or what researchers have called "design justice" [26]. Following the terminology of philosopher T. M. Scanlon, we propose a specific principle of design justice called "equal concern" [97]. The principle states that the way how AI systems are conceptualized, designed (and ultimately also deployed) should show due and equal concern to different stakeholders' interests. Equal concern prescribes that all affected stakeholders' interests should be taken equally into account when responsible agents are establishing what an AI system ought to be like in a broad sense—for example, what it should be used for and how. If the interests of certain stakeholders are not taken into account when answers to these questions are formulated, the organization responsible for the design process of an AI system fails to treat those stakeholders' interests as equally worthy of consideration, and thereby enacts what may be called a form of "design injustice".

Equal concern bears a close relation to the ideals of democracy and its close relatives, participation and inclusivity, which have been established as of utmost importance in the context of AI design and development. As philosopher of technology Mark Coeckelbergh has said, "if we endorse the ideal of democracy and if that concept includes inclusiveness and participation in decision-making about the future of our societies, then hearing the voice of stakeholders is not optional but ethically and politically required" [23]. Equal concern thereby calls responsible agents to implement design processes that are participatory and inclusive; that both involve the input of diverse stakeholder groups and are sensitive to the contextual, local and situated needs and interests any affected individuals and communities might have. Responsible agents should implement procedures and methods that allow affected individuals to voice their interests and concerns, respectively, and that effectively include them in significant decisions that affect their lives.

Importantly, showing equal concern towards stakeholders' interests does not mean that stakeholders actually have equally weighty interests or needs concerning the technology in question—or even legitimate interests at all. There can be legitimate and morally acceptable reasons for why—after first listening equally to each stakeholders' concerns and opinions—an organization developing an AI system might decide to prioritize certain stakeholders' interests. To provide a simple example, persons with disabilities or impairments certainly have a legitimate interest in having an AI system be equipped with an accessible interface. But to be reasonably accessible, an AI system does not have to have over 7000 language settings just because there are over 7000 languages in the world. In other words, equal concern requires that responsible agents seek to actively identify any legitimate interests that affected stakeholders might have, and they should also provide stakeholders the opportunity to actually voice their interests and concerns. Any identified interests and concerns should be treated as initially having equal weight, but it might be that, during a deliberative and dialogical process of stakeholder engagement and consideration, some interests and concerns turn out to have more weight than others. If this is the case, responsible agents should try to respect different stakeholders' interests in proportion to the weight of those interests.

Promote procedural, distributive and relational justice

The last prescription included under the principle of justice and fairness calls responsible agents to promote three kinds of justice in the context of their own work and conduct: procedural, distributive and relational justice.

- Procedural justice refers to the fairness procedures that are used to allocate benefits and burdens and to resolve disputes or conflicts. In the context of AI and automated decision-making, procedural justice requires things such as accuracy and integrity, non-discrimination and a lack of unjustifiable bias, transparency, public accountability and legitimacy, contestability, and access to remedy and redress.

- Distributive justice concerns the fair distribution of benefits and burdens, rights and basic goods, and resources and opportunities. In the context of AI and automated decision-making, distributive justice requires that individuals are treated as they are due in light of their merits, deservingness or need, for example, and that the resulting distribution of burdens and benefits is morally justifiable.

- Relational justice is about whether and how people can relate to each other as moral equals in societal and social contexts. In the context of AI and automated decision-making, relational equality calls for using technology in ways that repair past and present injustices, prevent marginalization and exclusion, and contribute to respectful relations between individuals and groups, for instance.

While the three are commonly distinguished from each other, they overlap in practice. For example, questions related to procedural and distributive justice in decision-making have to be considered together because the distribution resulting from an allocative procedure depends on the nature of the procedure.

From a procedural perspective, what should the algorithmic decision-making process be like for it to be fair towards individuals? At least three criteria of procedural justice can be distinguished here: First, the decision-making process should be accurate by way of being informed by facts and evidence and by being sensitive to individuals' claims to a given outcome—such as reasons for why they should receive a positive decision. This requires that the data and models used to make decisions for or about individuals are accurate and reliable, for instance. Second, the decision-making process should not be wrongfully biased in the sense that some individuals are treated comparatively better (or worse) without appropriate justification. This requires that the applied data and algorithms lack inappropriate forms of "algorithmic bias" that can lead to discrimination, for example. Third, individuals should be able to contest or challenge decisions concerning them and, when necessary, provided compensation or other access to other forms of redress [60]. This requires not only that mechanisms for contestation and redress are implemented, but also that sufficient transparency is ensured regarding the reasons and logic behind the decision-making process (see also [4]).

As a distinct form or domain of justice, distributive justice covers a complex set of overlapping considerations that relate to the question: how should things that we value be distributed among individuals in a just society? The relevant benefits in questions can include either (1) benefits allocated with AI systems, such as social benefits, access to opportunities and services, but also (2) access to the use of AI systems itself, for example, when considering access to assistive technologies. From the perspective of distributive justice, AI systems should be developed and used in ways that lead to fair distributions of benefits and burdens (whatever they may be) without discrimination on the basis of legally protected or otherwise sensitive attributes (such as gender, ethnicity, or sexual orientation). This means, quite practically, that the (expected) distribution goods or opportunities resulting from the use of the AI system should "track" whatever legitimate claims individuals have to the goods or opportunities in question. In medical contexts, those claims regard need and conditions of health, for example, whereas in making loan decisions one might be interested in income, financial stability and one's risk of defaulting on the loan. Importantly, we remind the reader that both non-discrimination law and moral obligations can require that AI systems should be proactively designed and used to enhance or improve the prospects of some individuals, such as members of marginalized or historically disadvantaged groups. In such cases, AI systems should be used to identify groups that lack access to necessary goods and services, for example, and to enact affirmative action policies in recruitment and educational contexts—provided that an appropriate legal and moral justification for such policies is given (see [114, 115]).

A last set of requirements from the perspective of justice in the context AI relate to so-called relational justice or relational equality. Relational egalitarianism is a philosophical conception of justice (or equality) motivated by the idea that a just society should not only distribute basic rights, goods and resources fairly among its citizens—it should enable people to relate to each other as equals in fundamental sense [5]. The principle of relational equality thereby underlines that justice is more than a matter of distributing goods: it is also a matter of social relations, democratic and political participation, and equal protection from the coercive use of public and private power, among other things. The moral requirement to work towards relational equality thereby calls responsible agents to develop and use AI systems in ways that foster and maintain respectful and equal relations between individuals and communities. On the one hand, responsible agents should refrain from supporting and reinforcing oppressive institutions, exclusionary social norms, and discriminatory societal structures, which should be abolished as opposed to upheld. In practice, this can mean that responsible agents ought to refuse from building or procuring AI systems that are (foreseeably) used to reproduce substantive inequalities, for example, or that are deployed to oppress minorities or marginalized groups. The presently discussed prescription relates closely to the previously discussed notion of equal concern in that promoting relational equality requires responsible agents to show special concern for marginalized communities and minorities, for instance. On the other hand, the notion of relational justice and equality has a more active and promotive component as well. In particular, responsible agents should actively search for ways to dismantle structural barriers that stand in the way of an equitable society—with and without technology. In other words, they should seek to disarm systemic and structural forms of oppression, which notably affect primarily groups that have historically faced discrimination and various types of disenfranchisement (see [26]).

Violation of laws, regulations or standards.

An AI system violates laws and regulations that apply to the system or the system's use-context, or the system does not adhere to the standards and norms of the use-context—including safety standards or professional codes of ethics, for example.

Lack of public justification

The operator of the AI system provides an inadequate public justification for using the system, fails to provide a legitimate and acceptable justification entirely, or does not otherwise meet standards of public accountability and transparency.

Lack of concern for stakeholders' interests

Affected stakeholders have not been identified and/or engaged during the conceptualization, design or development phase of the system, or relevant stakeholders' interests have been neglected, dismissed or afforded insufficient weight.

No recognition or compensation for participation

Persons who (knowingly or otherwise) participate in the design or development of the system are not recognized, respected, and fairly compensated. For example, their contribution (e.g., data work) is not recognized or is made invisible.

Lack of consideration for accessibility

The needs and interests of persons with disabilities are neglected, misrecognized or merely stipulated during the design phase. For example, the machine learning model or the interface have not been assessed for ableist biases.

Unlawful or wrongful discrimination due to bias

An AI system performs worse when used on members of a legally protected or otherwise socially salient group, its use leads systematically to worse outcomes for members of that group, or it is used primarily to harm members of that group.

Failure to fulfill positive duties

The goal, output or behavior of an AI system does not reflect the positive duties that the system operator has, such as the duty to provide reasonable accommodations, or the system prevents the fulfillment of such duties.

Non-comparatively wrongful treatment

An individual is not treated fairly in the use of an AI system. For example, the system fails to recognize an individuals' morally or otherwise significant merits, personal characteristics or achievements during an assessment of the individual.

Comparatively wrongful treatment

An individual is not treated fairly when the treatment in question is compared to how other individuals are treated. For example, members of some group are comparatively more likely to receive a false prediction or misclassification.

No mechanisms for transparency, contestation or redress

There are no means for users or decision subjects to access, evaluate, or contest the output of an AI system, or there is no way for individuals to seek redress and compensation when misclassified, mispredicted or harmed by the system.

Relational injustice

The AI system polarizes people, or contributes to the marginalization or social exclusion of some groups. Alternatively, the system sustains or exacerbates structural or substantive inequalities, or creates significant inequities in power.

Unfair (denial of) access to the system

Stakeholders who should have access to the system—such as members of key user demographics—are explicitly denied access to the system, or they cannot access the system due to excessive costs or other barriers.

3 REGULATING AI

Comments and recommendations on AI regulation and governance

REGULATING AI BASED ON RIGHTS

The political, legislative and regulatory level is the most important and effective level for conducting actions in service to the protection of fundamental values, rights, and freedoms. However, there is a wide debate on whether and how to regulation of AI systems and their use. This document recognizes these debates and argues that the appropriate approach is to center regulation on the protection of individuals' rights and fundamental freedoms. We maintain that the protection and active promotion of individuals' access to the rights and freedoms they are entitled to in democratic societies should be the key aim of technology regulation. Achieving this aim is not only valuable in itself—it is instrumentally valuable with regard to realizing fundamental values and for enabling and building a healthy, ethically oriented culture of innovation and business.

- ✓ Policymakers should improve (the enforcement of) existing legislation and regulations to protect individuals' human rights and fundamental freedoms from the effects of using AI systems.

- ✓ Policymakers should implement new rights for individuals, legal norms, and regulations concerning (the development and use of) AI to ensure the protection human rights and fundamental freedoms.

- ✓ Policymakers should implement bans and moratoria on (the use of) AI systems that are inherently in conflict with human rights and fundamental freedoms.

These recommendations will be detailed below together with more specific recommendations falling under each general recommendation, respectively. In line with existing proposals for rights-based approaches, our recommendations are grounded in the notion that individuals should be protected from unsafe, rights-violating, and ineffective systems; that they should not face discrimination by algorithms; that their data should not be manipulated or exploited; that they should have agency over the collection and use of their data; that they should be able to know and inquire whether, when and how an automated system is being used and how its outcomes impact them; and that they should have effective means of contestation and remedy when they encounter any problems in the context of AI system use. A central question, from the perspective of our framework, is not whether these rights and freedoms should be protected but, rather, what should be done at the regulatory level to ensure their protection.

Improving and enforcing existing laws and regulations to protect individuals' human rights and fundamental freedoms.

While AI systems can have some unique characteristics that more traditional forms of software lack—including, for example, high degrees of functional autonomy, adaptivity, modularity, malleability, and larger scales of use—most risks and effects that prove relevant for ethical and/or legal reasons are not unique or specific to AI technologies as such. This suggests that there is a need to refine and enforce existing laws and regulations that are relevant for preventing and addressing the various harms that AI systems—but also other technologies—can (be used to) enact. The rights and freedoms that individuals enjoy, and which have a well-established status in democratic societies, ought to be understood as concrete and effective tools to protect individuals from discrimination, violations of people's autonomy and dignity, and mental and physical harm, for instance, and the nature or application of a given technology should not change this fact.

In many cases, the improvement and more effective enforcement of already existing protections is a desirable and feasible way of addressing certain well-established risks involved in the use of AI systems in public and private contexts. For example, "algorithmic discrimination"—referring to acts of discrimination that are conducted with algorithmic systems, either intentionally or unintentionally—does not in many cases present significantly novel challenges from the perspective of legal protections against discrimination. However, it is clear that the algorithmic systems can be used to discriminate against individuals at larger scales and in opaque manners, and that there are also novel issues that should be considered and researched to determine whether they can be effectively addressed with existing protections against discrimination. These include, for example, affinity profiling, discrimination-by-association and discrimination against non-salient social groups [113]. Similar things might be said about many types of data protection violations resulting from mass-scale surveillance and data collection, where new types of problems are not necessarily introduced by the use of AI systems as such but, rather, existing problems are exacerbated and compounded by their ubiquitous and large-scale deployment.

It is also for these reasons that legal norms and regulations concerning non-discrimination, equality, and data protection (in particular) should be enforced better and more effectively, leveraging also the capabilities and capacities of the research community and NGOs to assist in this effort. Some norms, regulations and protections might need to be refined or expanded in order to better capture the specific problems associated with AI systems. Expansions in these respects should mandate responsible agents to follow standardized procedures and accountability mechanisms, for example, that work to ensure effective protection of individuals rights and freedoms. To counter algorithmic discrimination, for example, requirements for system developers or operators to conduct statistical tests or qualitative studies to identify problematic biases in AI systems as a part of equality impact assessment procedures should be implemented, and a requirement to conduct interventions on datasets or models to mitigate those biases should be established. It can also be desirable to formulate practical guidelines for assessing specific risks (e.g., algorithmic discrimination against socially non-salient groups) and to establish forms of collaboration between credited or overseeing bodies (e.g., auditors or ombudsmen) and industry agents.

New rights, legal norms, and regulations to ensure the protection human rights and fundamental freedoms.

Beyond refinement and better enforcement of existing laws and regulations, we suggest that implementing novel rights for individuals and laws and regulations concerning AI systems (or rather their development, distribution, and application) is necessary to ensure the protection of human rights and fundamental freedoms. They will have to be geared specifically for the context of AI and centered explicitly on the aim of protecting human rights and fundamental freedoms. Moreover, they should be defined and established with due regard for benefits of fostering a healthy and supportive climate for business and innovation. However, we emphasize that those benefits cannot and should not be achieved at the expense of individuals' rights and freedoms, or fundamental collective values that citizens of democratic societies accept. This document thereby joins many civil society organizations calling for a rights-based regulation of AI systems. In our proposal, a rights-based approach to regulation is envisioned to imply at least three sets of actions that should be undertaken by policymakers. These sets of actions are found below.

Rights of individuals. Policymakers should establish novel rights for individuals affected by the use of AI systems. These would include rights for people that are subjected to (semi-)automated decision-making such as a right to access evidence and a right to contest automated decisions.

Regulations and mechanisms for accountability. Policymakers should implement regulations concerning the use of AI systems, focusing in particular on systems that can or are expected to have significant effects on human rights and fundamental freedoms.

Bans and moratoria. Policymakers should impose categorical bans and moratoria on inherently or expectably rights-violating systems, such as on indiscriminate surveillance and biometric recognition in public spaces.

We will detail each of these recommendations below, suggesting also specific rights and policy mechanisms that we consider necessary to ensure that responsible agents comply with the demand of respecting individuals' rights and freedoms. Notably, our proposal for a rights-based approach differs in crucial respects from on-going efforts to regulate AI, however, such as the European Union's draft of the AI Act. In this light, we wish to provide the reader some background information on debates regarding the regulation of AI on the following pages, and to also motivate our proposal by enumerating some of the benefits of a rights-based approach in contrast to its alternatives.

A PERSPECTIVE:
Why regulate AI based on rights?

A rights-based approach to regulation stands in contrast to feature-based approaches, which seek to categorize AI systems based on specific technical or functional features, and risk-based approaches, which aim to regulate the use of AI systems based on their expected effects. An example of the latter is found in the European Union's draft regulation concerning AI which classifies AI systems into four categories: (1) systems with "unacceptable risk" that are to be banned, (2) systems with "high risk" that are to be subject to stricter demands (e.g., risk assessment and mitigation systems and documentation requirements), (3) systems with "limited risk" for which there are specific transparency obligations, and (4) systems with "minimal risk" that are subject to no specific requirements. We posit that a rights-based approach has clear advantages, many of which are noted previously by research, non-profit, and civil society organizations such as AccessNow.

Against an feature-based approach, we note that "artificial intelligence" is a moving target that escapes definition. Regulation that proceeds by characterizing the features of "AI systems" (e.g., adaptivity or autonomy) will likely end up chasing a definition for an effectively fluid and moving target. In many cases, such as with decision tree methods, the difference between "AI" and traditional computation is also one of scale or degree, and there is no strict difference to "traditional software". A rights-based approach to regulation avoids this issue, and also accommodates for the fact that the relevant harms that individuals and communities should be protected from can also be generated through the use of more traditional software (e.g., rule-based and manually coded systems). While a risk-based approach is preferrable to feature-based one, such an approach can also fail to establish the necessary protection needed to safeguard individual rights. This is in part due to the transportable, modular, and malleable nature of AI systems which creates risks for dual-use and means a given system's level of risk can vary abruptly. Due to contingent factors that manifest through an AI system's use-phases or due to (dual-)use of the system for malicious purposes, an initially trouble-free system can be quickly rendered highly dangerous and unsafe. This is especially the case with online learning systems and regularly updated systems, which are affected by changes in the use-context and the affected population. As it has been frequently noted in debates about regulating AI, a "high-risk" classification in this sense does not provide sufficient moral and legal reasons to warrant the deployment of any given system.

In our view, protection of individuals' rights and freedoms should not depend on perceptions of risk or the nature of the technology in question. (This is true even if we acknowledge that practical reality may throw us into morally tragic situations where it is impossible to ensure that everyone can access and exercise their rights and freedoms.) A rights-based regulation avoids the abovementioned problems by ensuring that affected individuals are protected by law and guaranteed to have access to their rights and freedoms. A rights-based approach can better accommodate for the fact that morally and legally relevant reasons for allowing, omitting or ceasing the use of a given system may arise through time due to contingent factors. Thereby, the rights-based approach also has independent grounds for requiring at least some forms of oversight even when a given AI system does not present immediate risks and expectable issues. Such an approach suggests the necessity of accountability mechanisms—such as regular and iterative impact monitoring processes concerning algorithmic systems—in all or most cases.

Still, adopting a rights-based approach to regulation does not mean that certain systems bear more significant risks or that assessments of risks and impact is pointless or redundant from a regulatory perspective (or indeed prior to the deployment of an AI system). Risk assessments can indicate a need to prioritize efforts to protect individual rights in specific public and private contexts, sectors or domains (e.g., public spaces and other contexts which we have termed "high-stakes decision-making contexts"). Public decision-making contexts can be understood as priorities, for example, due to the nature and significance they bear as areas and mediators of human, social, and political life. A risk-based approaches can in principle thereby be complementary to a rights-based approach, but this should not mean that the notion of "risk" is centered from a legal or moral perspective. Risk management activities can and should complement efforts to safeguard individuals' rights in specific, local contexts of technology development and use. Given the malleable nature of AI systems and the context-dependent nature of their impact, assessments of risk are best conducted locally—by parties that are best situated to assess risks that may arise with respect to specific systems that are under development or already in use.

Automated decision-making systems can make simple mistakes with detrimental consequences, and their predictions and other outputs can involve unacceptable biases that can in worst cases lead to discrimination and unfair treatment. Yet their complexity can prevent decision subjects and other agents from accessing information about why and how the output was generated. In line with the spirit of Recital 71 in the GDPR and the revisions made by the European Commission to the Product Liability Directive, we recommend the introduction of an individual right of access to evidence to mitigate these issues. An individual right of access to evidence regards (a) evidence of an erroneous or otherwise significantly harmful output or prediction used for the purpose of making a high-stakes decision, including also (b) the data used for generating the output or prediction and (c) information regarding the nature of the inferences and processing involved. Natural persons should be able to invoke this right during legal proceedings to access the abovementioned evidence from the party operating an AI system (including, for example, companies and suppliers). This right would function to protect core values such as justice and respect for human autonomy by ensuring sufficient transparency and establishing an effective way for victims to seek legal reparation, redress, and compensation in case they are mispredicted, misclassified or otherwise harmed in the use of AI or automated systems in high-stakes contexts.

A right to contest (semi-)automated decisions. Policymakers should implement an individual right to contest (semi-)automatically generated decisions in high-stakes decision-making or operational contexts.

We propose that policymakers should implement a right to contest (semi-)automatically generated decisions. A right to contest refers to an affected individual's right to challenge decisions made or supported by AI systems (or other automated systems). As legal theorist Margot Kaminski notes, "a right to contest AI is both normatively desirable and practically feasible" and it "could ameliorate the foreseeable harms of AI" [60]. While contestation is not sufficient for ensuring justice in decision-making procedures, it is necessary for this purpose. Contestation is a historically well-established adversarial mechanism employed for the central purpose of preserving and enhancing procedural justice in democratic societies. We might also note that a right to contest AI is grounded in the notion of due process and bears also a close relation to the right to the contest automated decisions as established in article 22(3) the General Data Protection Regulation (GDPR) which states that "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision".

A way to regulate "algorithmic shadows". Policymakers should implement a reasonable policy mechanism for handling unlawful "algorithmic shadows" left on machine learning models.

Machine learning models trained on individuals' personal data include a "persistent imprint of the data that has been fed into a machine learning model and used to refine that machine learning system" [69]. Algorithmic shadows constitute an area of concern for various moral and legal reasons.

- First, the presence of algorithmic shadows in machine learning models can expose individuals to privacy violations, for example, because models can be vulnerable to attacks—such as model inversion attacks and membership inference attacks—that can be used to infer sensitive information from the model. For this reason some researchers have argued that certain machine learning models themselves could be legally classified as personal data according to the GDPR [110].

- Second, while various privacy-preserving methods and tools are motivated by this issue in that they seek to prevent malicious attackers from inferring personal or otherwise sensitive data from the model, these methods do not address a related problem—namely, that the data used to train models can include data that is collected from individuals in an unlawful or exploitative manner. In other words, algorithmic shadows can be the result of exploitative data practices which should be disincentivized and penalized for independent reasons. Arguably, non-compliant agents should not be able to profit from the fact that they have access to higher performance models by engaging in illicit data collection practices, for example.

- Lastly, given that parties operating AI systems that embed algorithmic shadows can use those models for malicious purposes, erasure of algorithmic shadows can be required to prevent concerned individuals from being made unwillingly complicit to individual and social harms enacted with the system.

We recommend implementing a reasonable and proportionate policy mechanism to address algorithmic shadows, especially in cases where such shadows result from an agent engaging in illicit practices of data collection or transfer. Data deletion rights—such as the right to erasure established in article 17 of the GDPR—can ideally contribute to mitigating this issue, but data erasure is largely insufficient to address the problem "due to the presence of algorithmic shadows that persist even after data are deleted" and which "can still cause harm to an individual and to groups that relate to that individual" [69]. There are other options, including classifying certain machine learning models as personal data [110] or implementing an individual right against unlawfully imprinted algorithmic shadows. However, we also note that a proportional and effective mechanism could be an enforceable penalty of "algorithmic destruction". This mechanism has been previously used by the U.S. Federal Trade Commission to mandate the destruction of companies' models and algorithms that have been trained on data collected from individuals without their meaningful consent.

The previous recommendations provide examples of what a rights-based approach should look like in practice. However, policymakers should also implement a further set of policy mechanisms and regulations to ensure that the development or use of AI systems does not violate individuals' human rights and fundamental freedoms. These include, for example, data governance controls and risk management procedures that should be established within organizations to ensure compliance throughout an AI system's lifecycle. For now, we note that two crucial points should be kept in mind when designing applicable regulations and mechanisms that serve to promote compliance and "algorithmic accountability":

- First, certain agents—most notably, public agents of specific sectors—can have positive legal duties to actively promote individuals' access to rights and freedoms, a prominent example being the positive duty to enhance substantive equality and to provide reasonable accommodations to people with disabilities. Any policy mechanisms or regulatory approach seeking to protect individuals' rights and freedoms should account for the necessity of fulfilling positive duties, including by determining appropriate standards for compliance with those duties when (semi-)automated decision-making systems are used to carry out decision-making activities (e.g., recruitment or university admissions).

- Second, as is correctly assumed in the risk-based approach to regulating AI, technologies and systems that are used in specific contexts—in particular, ones that involve "high stakes" or significant risks from the perspective of individuals rights and freedoms—ought to be subject to stricter requirements than other systems. There is a need to prioritize certain contexts, in other words. However, as AI systems are often designed to be highly malleable, modular, and scalable, implementing appropriate regulations and policy mechanisms that anticipate risks for dual-use (and changing purposes of use) is necessary. Any established regulations and protections—and any mechanisms that seek to promote and ensure accountability on part of system developers or operators, more generally—should be designed for purpose limitation with respect to the objective or purpose of using a given AI system, and to anticipate cases where systems or algorithms travel across hands and use-contexts—for instance, when systems designed for one geographical location or task end up later used in another location or task.

In Section 6, we describe desirable and feasible practices and mechanisms that can and should be implemented to protect individuals' rights and freedoms, and to ensure that AI systems are used with due regard for public accountability and the rule of law. Ideally, the practices and mechanisms we discuss would be mandated by law, but we note that responsible technology developers and operators should in any case implement them as a means of voluntary self-regulation. We reserve the discussion of these practices and mechanisms for another section for this reason, and because the discussion will be more detailed by virtue of digging more closely into topics such as algorithmic impact assessment, dataset and model documentation, and explainability in AI. Importantly, many of the practices we discuss are already included—at least in some form—in other proposals for AI governance and regulation, including the AI Act draft. Our recommendations are meant to be complementary in this regard.

Banning AI systems and applications that are incompatible with human rights and fundamental freedoms.

This document joins in recent calls for prohibitions on certain uses and applications of AI systems to be implemented at the level of (inter)national laws and regulations—namely, any and all AI systems that necessarily or foreseeably stand in conflict with human rights and fundamental freedoms. Prohibitions should obtain at least until the emergence of such conflicts can be effectively prevented. We will next provide a list of prominent candidates for categorical bans and prohibitions. Our list of proposals includes prohibitions that are identical or similar to prohibitions that have been previously proposed in the draft of the AI Act or by other entities such as AccessNow. We have also taken into account a number of opinions and statements that have been presented in discussions around on-going regulatory processes, such as those made by the European Data Protection Supervisor and the European Data Protection Board.

The recommended set of bans is found on the following page. The list is not exhaustive, but rather indicative of the types of applications that can be considered incompatible with respect for fundamental rights, freedoms, and core values of democratic societies. Importantly, even if the proposed list of bans and prohibitions is not implemented at a national or international level, we recommend that responsible developers and operators of AI systems refrain from developing or using technologies for the purposes described in the previous list. We also recommend that responsible agents should also establish any other "red lines" or prohibited use-cases that they deem necessary and appropriate in their specific context of operations. This is because responsible technology developers and operators—including organizations such as companies and non-profits—can be suitably positioned to make localized and contextual judgments regarding what ought not be built or deployed in a given social or political context.

☒ A ban on fully autonomous weapon systems

A ban on fully autonomous weapon systems, where 'fully autonomous' refers to systems where no direct and real-time human control is exercised over the selection of targets, the triggering of defensive or offensive measures or any other measures that may inflict physical harm to individuals.

☒ A ban on indiscriminate and/or arbitrarily targeted processing of biometrics

A ban on indiscriminate and/or arbitrarily targeted use of systems for automated recognition, identification, categorization, authentication, or verification of human features (such as emotional states, faces, gait, DNA, voice, keystrokes, and other behavioral signals, affect, gender identity, ethnicity, intent, political or sexual orientation) in publicly accessible spaces (both physical and online) regardless of temporality (both real-time and post data capture).

This prohibition excludes the use of AI systems for the detection of human presence in public spaces, provided that (a) there is a lawful basis for processing and (b) no subsequent changes made to the objective function or use-purpose of the system.

☒ A ban on autonomy-bypassing and exploitative systems

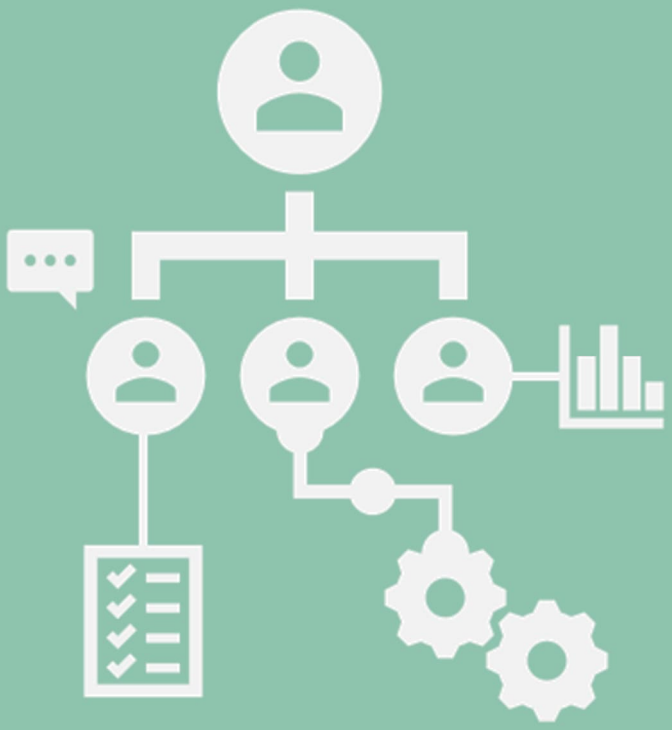
A ban on the use of systems that (1) employ hidden influences beyond a person's consciousness in order to influence their reasoning or materially distort their behavior, or (2) which exploit any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, or any other grounds on which discrimination is prohibited for legal or moral reasons. (See article 5 of the draft AI Act.)

☒ A ban on "social scoring"

A ban on information processing systems developed to calculate or establish a "social score" or a relevantly similar output concerning individual persons or groups. The envisioned ban should apply to (1) public and high-stakes decision-making contexts, where those calculations result from the evaluation or classification of natural persons based on their (a) physical attributes, (b) social behavior, or (c) known or predicted personal characteristics, and (2) other semi-public and private contexts where the use of such information cannot be demonstrated to be compatible with human rights.

5 AI ETHICS AT THE ORGANIZATIONAL LEVEL

What should organizations do?



WHAT SHOULD ORGANIZATIONS DO?

To both ensure compliance and build a pro-ethical culture across contexts of AI system development and use, ethical objectives and practices need to also be effectively implemented at the organizational level. This document maintains that the organizational perspective is crucial for the effective implementation of the lifecycle ethics approach to the design, development, and use of AI. Throughout this section we wish to convey the message that a collective commitment to ethics should be tangible and visible in day-to-day practices and it should manifest also at the level of the roles, structures, and norms of behavior within an organization. Those structures, roles, and activities—among other factors, such as organizational culture and climate—condition whether and how values are adopted and integrated into real-life practices of technology development and use [121]. They also affect the execution and effectiveness of pro-ethical design and development practices. After all, ethical thinking and moral behavior is situated, local, and contextual, and it takes place amidst concrete organizational practices and work environments which have their respective affordances and limitations that affect and structure people's work, attitudes, and their ethical deliberation. In a nutshell, our recommendations concerning the integration of ethics at the organizational level include four sets of actions, which are also summarized in the table found on the following page.

Awareness. Responsible agents should increase awareness of ethics within organizations designing, developing, or using AI systems.

Culture. Responsible agents should build a common culture of responsibility, lawfulness, and ethics in their organizations.

Practices, roles and structures. Responsible agents should introduce norms, roles, responsibilities, and practices around ethics within their organization.

Skills and tools. Responsible agents should develop and support others' development of skills, mindsets, and competencies.

The overarching lesson here is that ethics in the context of AI is not about mere compliance, nor is it simply about building machines that "align with human values". Rather, ethics is an on-going and ever developing orientation towards care and responsibility that is integrated into—and enabled by—a broader set of structures and norms within organizations, as well as active efforts to uphold a lawful and ethical culture therein.

RECOMMENDATIONS:
Integrating ethics at the organizational level

STEP 1. Increasing awareness about ethics and social responsibility

- **Educational resources and/or training.** People are provided access to educational resources or events and/or internal or external training regarding ethical issues relevant to their work.
- **(In)formal activities around ethics.** People within the organization organize formal and/or informal activities around ethics, such as reading circles or writing blog posts.
- **Certification systems.** People within the organization can establish and certify skills regarding AI ethics within or outside the organization, such as by completing online courses for pro-ethical design.

STEP 2. Building an organizational culture of responsibility and accountability

- **"Organizational nudges".** There are slogans, actions, reminders, and rubrics that prompt ethical reflection and consideration in people's day-to-day decisions and work tasks, and which structure people's behavior by reminding them to take the ethical implications of their own actions and decisions into consideration.
- **People interpreting their work in ethical terms.** People interpret their own actions and work in ethical terms, possibly by using a shared language established for discussing ethical issues within the organization.
- **Pro-ethical behavior is protected and incentivized.** Legal protections for whistleblowers are respected in the organization. Employees are free to engage in ethical trouble-shooting and they are incentivized to voice concerns about ethical issues that pertain to developed or used AI systems, components, products and services.

STEP 3. Integrating ethics into organizational practices, structures, and norms

- **Well-defined practices, standards and requirements.** There are well-defined ethics practices regarding ethical design, development and use of AI systems. Clear time-frames, requirements, instructions and targets are defined for those practices, including also documentation requirements.
- **An ethics team or an advisory board.** A person or team is officially in charge of ethical aspects of design, development and use of AI systems. If such persons are not available, responsible agents might wish to consult an external advisory board with the necessary forms of expertise.
- **Regularized practices which allow for deliberative flexibility.** Employees understand what is expected from them with regard to ethical aspects of design, development and use of the AI system. They have a standard procedure that they can follow, but they also know whether and when they can defer to their own contextual judgment.

STEP 4. Having the right people for the right job with the right skills and tools

- **Broad expertise for a broad set of tasks.** The person or team in charge of ethics possess expertise in (technology) ethics and social sciences and skills in communication, stakeholder management, analysis, and problem-solving. The person or team is equipped to handle their central tasks.
- **Diverse pro-ethical practices.** The person or team in charge of ethics conducts a variety of tasks regarding ethics. For example, they provide consultation, research (approaches to) ethical problems, create spaces for discussion about ethics within and outside the organization, empower other people to evaluate and refine systems in terms of ethical aspects, communicate ethical issues within and between teams, and identify stakeholders and manage relations with them.
- **Careful use of effective tools.** The person or team in charge of ethics has access to a variety of tools and methods that are necessary to operationalize ethical values and to refine developed or operated systems. They know the benefits and limitations of applied tools and methods, and are able to apply them effectively and safely.

In this section, we will look at what organizations aiming to integrate pro-ethical and responsible AI practices should do in order to ensure that integration is successful and sustainable. As it can be valuable to learn from failures, we will first take a look at what pitfalls they should avoid. In particular, we describe four pitfalls that organizations might fall into in the context of integrating ethics into practice (Table below).

TABLE. Four pitfalls at the organizational level of AI ethics

☒ **There is a lack of awareness about ethics in the organization**

- The organization lacks knowledge about ethics (e.g., why and how ethics is important in their case)
- The organization has misperceptions concerning ethics (e.g., ethics is perceived as constraining innovation or profitability)
- The organization fails to identify the benefits of ethics (e.g., intrinsically valuable but also legal and reputational benefits)

☒ **There is a lack of (perceived) responsibility in the organization**

- The organization lacks a culture of accountability and/or knowledge of its own responsibilities (e.g., compliance)
- Responsibilities are unknown or badly distributed within the organization (e.g., due to complexity and "many hands")
- Unethical behavior results from prioritization of business objectives over legal and ethical norms (e.g., ethics-washing)

☒ **The organization lacks explicit and well-defined roles, responsibilities, and practices around ethics**

- The organization lacks clear ethical objectives (e.g., well-defined standards or targets)
- The organization lacks incentives and safeguards for pro-ethical behavior (e.g., protections for whistleblowers)
- The organization lacks concrete practices, roles, and responsibilities centered on ethics (e.g., ethics board or team)

☒ **The organization lacks the right people with the right skills and tools**

- Pro-ethical practices center on the technology as opposed to broader organizational conduct (e.g., business model)
- Ethics teams lack key skills and diverse perspectives (e.g., expertise, socio-cultural diversity)
- Ethics teams lack tools and/or knowledge of how to use them effectively and safely (e.g., limitations or externalities)

We will discuss these pitfalls and problems in a bit more detail below. Our examination of these pitfalls also provides background for our corresponding recommendations on integrating ethics at the organizational level. Our recommendations are based on prior research and reports [67, 121] and they can be viewed as recommendations to establish organizational capacities corresponding to four levels of "AI ethics maturity", as it were. However, we emphasize that our proposal and recommendations should not be understood as a maturity model in the strict sense.

☒ There is a lack of awareness about ethics in the organization

A first set of problems relates to the simple fact that awareness is required for organizations to even begin to think about integrating pro-ethical perspectives and activities into practice. A company, for example, might not simply be informed about why and how ethics relates to their industry, or business model, operations. They may be unaware how ethical concerns might relate to the technology they develop, or the specific AI system they are using, or to a given use -case they are inquired about. They might not know, in other words, why it needs to meet certain standards or what ethical questions relate to the technology in question. A lack of awareness might have many reasons: It might be due to the fact that engineering education has not traditionally involved a sufficient ethics curriculum, or it might be due to the ethics playing a dishearteningly marginal role in business overall. But there might be also misperceptions concerning ethics. For example, business ethics or technology ethics might be incorrectly viewed as primarily restricting innovation and creativity, or as standing in the way of generating utility and profit. Even in the absence of misperceptions, an organization might simply also fail to identify the benefits of ethics in technology development and use—it might just be viewed as incurring further costs, for example.

☒ There is a lack of (perceived) responsibility in the organization

A second set of problems can arise even when the level of awareness about ethics is high in an organization. People within the organization may—even if highly conscious of ethical issues—lack knowledge about what is required from them specifically in terms of compliance and social responsibility, for instance. In other words, even if knowledgeable about the importance of ethical conduct, an organization using an AI system might not know what they should do exactly in the sense that there is uncertainty about concrete responsibilities. This problem related to being uninformed about compliance requirements can be exacerbated by the fact that software development pipelines and supply chains can be highly complex. Complexity introduces risks for miscommunication between different parties, such as developers and clients or management and regulators, for example. However, it may also introduce the so-called "problem of many hands" which refers to situations where different agents' specific responsibilities become blurred by virtue of there being so many agents involved. There might be many large teams working on a given data science project or an machine learning –based application, for example, which can sometimes result in ambiguity regarding each person's specific responsibilities.

Organizations might also fail to fulfill their respective obligations due to a lack of sanctions for noncompliance, or because they perceive their burdens as unfair. Unethical behavior may also be due to organizations consciously placing business objectives as a priority above ethics. This type of misconduct occurs when organizations engage in "ethics-washing" by giving the public the impression that they are behaving responsibly without actually doing so [38]. They might only "cherry pick" values that align with their business interests, for instance, or revise their ethical commitments to justify their pre-existing practices, allowing them to avoid changing unethical business models. Part of the problem here is that the organization lacks a culture of accountability and a genuine commitment to public values.

☒ The organization lacks explicit and well-defined roles, responsibilities, and practices around ethics

Even accountable organizations that mindful of ethics might have no explicit and well-defined ethical objectives and practices that concretely guide and constrain individual workers' behavior. There might be no clear vision as to how ethics should be implemented at the strategic and operational level, for example. If such a vision is lacking, ethical troubleshooting at the practical level can remain erratic or inconsistent because employees do not know what to search for and how. The organization might also lack incentive structures and safeguards that are necessary for sustaining ethical behavior among the workforce. Even though ethical troubleshooting might be appreciated, for example, this and other types of pro-ethical behavior might remain dependent on individual persons' motivation (which is undesirable). Or it might be that ethical concerns simply remain unvoiced, or that they might be unintentionally downplayed by certain people in management when expressed by employees. Crucially, the organization might also lack well-defined and explicit practices, roles, and responsibilities centered on ethics. For ethics to become effectively integrated into day-to-day activities and practices, there should be roles that involve relevant responsibilities that concern the ethical aspects of the developed or operated technology, for example, as well as standardized and regular practices that people in those roles should engage in. Otherwise ethical considerations can remain informal, occasional, and unstructured as there are no well-integrated and explicit protocols and accountability structures. This can be problematic as ethical considerations and practices might—unless explicitly built to include a standardized but holistic processes and reflective exercises— be inconsistent or remain narrowly bound to specific perspectives.

☒ The organization lacks the right people with the right skills and tools

Even when there are well-integrated pro-ethical objectives and practices in place, a lack of skills and tools that are necessary for purposes of solving ethical and legal problems can prove problematic in many cases. For example, teams conducting ethical review might lack relevant perspectives from social sciences and the humanities with the result that ethical considerations might tend to remain primarily centered on "broken parts" of the technological system—such as the training data or algorithms. This means that the fundamental ethical issues that underlie the identified problems—namely, the problems of which the "broken parts" are only symptoms—can remain unaddressed. When assessed from a broader perspective and with a more diverse set of conceptual lenses, however, the identified issue might be in fact traceable to the business model of the company in question, or to underlying rationale of an institutional procedure which the AI system is envisioned to optimize. The core problem in these cases of misidentification is that the organization lacks employees who possess the key expertise, attributes and skills required for holistic pro-ethical evaluation and design—such as knowledge and expertise of the system's use-context—and consequently remains too homogeneous to account for certain necessary socio-cultural perspectives. Crucially, even a diverse and knowledgeable team might nonetheless lack appropriate tools, techniques and methods for operationalizing data ethics and AI ethics in their own work. Alternatively, they might have many applicable tools without knowledge of how to use them. Even worse, they might use them without knowledge of their limitations. This suggests also knowledge of applicable tools—including how to use them—is required for effective pro-ethical AI practices.

INTEGRATING ETHICS: FOUR RECOMMENDATIONS

The previously described pitfalls can be understood as corresponding to failures on four levels of "AI ethics maturity", where the term in question is understood broadly and in a non-technical sense. We will next consider how to avoid those pitfalls and how to integrate AI ethics effectively into organizational practices and structures, respectively. We offer four general recommendations that are considered individually.

Increase awareness. Responsible agents should increase awareness of ethics within organizations designing, developing, or using AI systems.

To start building an approach to pro-ethical AI at an organizational level, awareness about ethics as it relates to the context of technology development and use should first be raised within organizations. Awareness-raising does not mean simply holding a meeting about the organization's ethical commitments and vision, but addressing misconceptions and misperceptions about the role ethics plays in technology development and use, for example, and in business more generally. There are at least three general ways to increase awareness, which might prove useful:

- Providing access to online courses and learning opportunities can help raise awareness, and organizing internal ethics training can enable employees to develop capacities for both identifying and addressing ethical concerns specific to their work. For example, a concrete way of doing this is to provide people access to online courses on AI ethics (such as the AI Ethics MOOC by Helsinki University) or encourage them to participate in workshops, conferences, or other events focused on the topic.

- From reading circles and blogs to research projects and promotional events, both formal and informal activities organized within and outside the organizations can help raise awareness about ethics in the context of AI. These activities are easy to implement and require little resources. Meanwhile, they can encourage self-development and communication between different employees and teams.

- Certification systems can be used to both establish internal ethical standards for AI design, development, and use, and to raise awareness about general and specific ethical aspects related to AI. Notably, certificates are also sometimes awarded as a result of completing online courses organized by universities and education centers, which means the establishment of certification systems can go hand-in-hand with offering employees access to educational resources and training.

Build a culture of responsibility. Responsible agents should build a common culture of responsibility, lawfulness, and ethics in their organizations and in the industry overall.

The previously mentioned ways of increasing awareness also support the second objective of fostering individual and collective responsibility within and across organizations. In practice, building a culture of responsibility, lawfulness, and ethics means that the organization should take steps to design working environments, behavioral norms, and day-to-day practices within the organization so that people's attention is actively directed towards ethical considerations regarding the developed or used AI products and services, including their social impact. What would be some desirable and feasible ways of achieving this?

- First, one could implement "organizational nudges" that help people think and behave with due regard for moral norms and ethical values. Organizational nudges can include slogans, actions, reminders, and rubrics that prompt and structure ethical consideration and decisions in day-to-day practices, for example. These kinds of nudges serve to direct people's attention towards ethical aspects of design, development, and use, and to structure behavior so as to ensure that people who work on their own specific tasks will take the ethical implications of their actions and decisions into consideration. For example, creating a shared language through mission statements, acronyms and slogans, and utilizing reminders or checklists, can help maintain people's focus on ethics in their day-to-day work.

- Second, and this is related to the previous point, one might encourage people to interpret their actions and work in ethical terms. Integrating ethics into work tasks and organizational practices means that people should actively interpret their own doings from the perspective of ethics and social responsibility. To assist in this, organizations should make room for discussions about values and ethics in formal and informal contexts. If the organizations has established a shared language for discussing about ethical issues, using that shared language in task descriptions and summaries—such as in task or system documentation templates—can be helpful.

- Lastly, building an organizational culture of responsibility requires actively protecting and safeguarding pro-ethical behavior. Recent developments in Big Tech have shown the importance of ensuring that practitioners and researchers can voice their concerns and intervene on ethical violations without fearing retaliation [43]. Everyone within the organization should be free to engage in ethical trouble-shooting and to voice concerns about ethical issues. Legal protection of whistleblowers is necessary in this respect. From an organizational perspective, furthermore, there is also a need to appoint a person—such as a shop steward—or an internal or external board to ensure that employees have someone to turn to when they wish to voice concerns about ethical issues in the organization.

Integrate ethics into organizational practices, structures and norms. Responsible agents should introduce norms, roles, responsibilities, and practices around ethics within their organization.

An organization-wide commitment to ethics should be visible in day-to-day practices within the organization—careful ethical reflection, troubleshooting and problem-solving should become part of the "business-as-usual". In other words, a compliant and careful approach to central activities and key tasks should become the way things are simply done by default, and the way people also expect things to be done. But what would different kinds of pro-ethical practices look like exactly? regular ethics training, research activities, ethical review processes, algorithm audits, and stakeholder engagement activities. They might also include tasks such as providing consultation, researching ethical problems related to AI systems and approaches to those problems, creating spaces for discussion about ethics within and outside the organization, empowering other personnel such as model builders to evaluate and refine systems in terms of ethical aspects, communicating ethical issues within and between teams, identifying stakeholders, and managing relations with them. In addition to having such practices introduced within the organization, those practices should also be effective in achieving their constitutive aims. Below we describe three sets of considerations that should be taken into account when integrating pro-ethical practices [107].

1 Pro-ethical AI practices should be well-integrated.

Ethical reflection and attention to legal and moral norms are perceived as a natural part of central tasks and processes throughout the AI system's lifecycle. Self-initiated ethical reflection and ethics-related tasks that people are required to complete, for example, are not perceived as merely superficial, extraneous or restrictive. Ethics is rather understood as valuable from a procedural perspective of due diligence and as central to achieving successful outcomes.

2 Pro-ethical AI practices should be explicit.

Practices and requirements related to legal and moral aspects of technology are well-known within the organization. They are also articulated as central to each relevant role, process, and project. When pro-ethical practices are explicit, the tasks that people conduct to ensure adherence to moral and legal norms do not remain merely implicit—they are not conducted informally in silence, for example—and people are not able to simply forget or overlook their tasks and responsibilities.

3 Pro-ethical AI practices should be regularized.

There are clear structures and requirements for ethics-related tasks and procedures. Practitioners should be able to develop their capacities and strengths in ethical analysis and decision-making through repetition and regular engagement in relevant tasks. When regularized, ethics-related practices do not remain unstructured—for example, their execution, success or effectiveness does not vary significantly depending on who conducts the task. Persons working on ethical aspects of AI products and services do not have to defer to their own views and opinions in each case.

Right people with the right skills and tools. Responsible agents should develop and support others' development of skills, mindsets, and competencies required for pro-ethical technology design, development, and use.

Establishing clear roles centered on ethics is central to ensure the effectiveness of the various pro-ethical AI practices that were described previously, and we have emphasized that those roles should involve explicit and well-defined responsibilities. However, at an organizational level, it should be also made sure that people who are in the relevant positions have the rights skills and tools for the job. But what kinds of skills are relevant? Pro-ethical practices require knowledge of concepts, theoretical lenses, and modes of moral reasoning. This poses the need for training in philosophical ethics, including technology ethics and domain-specific ethics relevant to the system use-context. Furthermore, as technologies operate within social and political contexts, expertise in social sciences—including also human-computer interaction, for example—and humanities is also increasingly necessary for assessing the implications of technology deployment. In practice, AI ethicists also need to analyze and synthesize various forms of (non-)technical information from different domains in order to identify problems of a sociotechnical nature. They need to also communicate complex ethical issues across stakeholders and audiences with varying amounts of technical knowledge and expertise. Given the various stakeholders involved and the speed of software development processes, those persons also need to facilitate decision-making under complexity and uncertainty (recognizing also that there can be reasonable disagreement about business and ethical values between different stakeholders). Consequently, there are also various analytical skills, problem-solving skills and communication skills that responsible practitioners require.

To be effective in their tasks, persons or teams in charge of ethical aspects of design, development and use of AI systems ought to have access to applicable tools and methods. For example, improving model fairness or extracting explanations for predictions tends to require the use of technical toolkits and state-of-the-art methods. There are many types of different tools and methods available for various different purposes, nonetheless, which is why responsible agents should know what tools and methods to use (and when). We have listed some general categories different pro-ethical AI tools and methods in Table 5 on the following page, including also their respective benefits and limitations from the point of view of implementation in the software pipeline. Within each category, one can find also different tools with their respective benefits and limitations. We encourage responsible agents to research and identify those benefits and limitations to ensure the safe and effective use each tool. For example, we might note that there are dozens of metrics for fairness which can be used to evaluate models [111] and tens of methods to extract explanations [81]. Yet overreliance on metrics, for example can be problematic because metrics tend to transform into targets that are optimized for their own sake [106]. Machine learning system developers should use diverse sets of metrics, accordingly, and combine quantitative metrics with qualitative methods for generating information in order to get a holistic picture of the impacts of their models (and systems overall). The key point here is that, before applying pro-ethical AI tools in practice, responsible agents should know whether, how, and when they should be applied.

PRO-ETHICAL AI PRACTICES:
What should be in the ethicist's toolkit?

Ethics frameworks and other theoretical resources. Philosophical theories, human rights frameworks, ethics frameworks for professional contexts/domains, industry standards, and codes of conduct

☑ Highly available and can help structure practitioners' ethical reflection (e.g., what are salient risks in a given contexts, what "count" as ethical issues in that context).

☒ Use of theoretical resources requires contextualization (i.e., considering what a theory would prescribe or recommend in a specific case or in a local context of technology use).

Checklists, rubrics and guides. Documents or heuristics designed for specific tasks (e.g., data collection, model evaluation, UX design) and for realizing specific values (e.g., privacy, fairness, transparency).

☑ Highly available and can be used to structure tasks and practitioners' ethical deliberation. Require little to no resources to implement and can be particularly useful in well-specified tasks (or during particular stages of the machine learning pipeline such as model evaluation).

☒ May incentivize a "box-ticking" mentality and use may require supplementary material and/or contextualization.

Metrics and technical toolkits. Evaluation metrics for identifying and assessing system impact (e.g., fairness metrics) and technical toolkits for improving the system (e.g., explanation extraction, bias mitigation)

☑ Highly available for different purposes (e.g., bias testing, database and model privacy estimation, and explaining machine learning models). Implementation requires little to no resources.

☒ Typically available for specific tasks or stages of the system pipeline (e.g., model fairness). Effective and safe use requires expertise due to lack of existing standards, detailed guidelines and best practices

Research and case-studies. Research outputs that describe and analyze areas of concern (e.g., system transparency) and provide ways of improving systems or processes.

☑ Can assist in the identification of risks and solutions, and provides insight into well-known issues.

☒ Often not much research on marginal use-cases or domains. Drawing insights from research outputs can also require technical expertise (or other forms of domain-expertise).

Channels for diverse input and feedback. Methods for gathering information outside the team or organization (e.g., user feedback, stakeholder engagement, surveys).

☑ Can provide developers or operators localized and contextual insights and information, and channels can be tailored to gather information about specific themes or areas of interest.

☒ May contain "noise" or bias towards negative aspects (e.g., primarily negative user feedback). Resource-intensity depends on channel (e.g., user surveys versus collaborative design).

5 TAKING ETHICS INTO PRACTICE

What should technology developers and practitioners do?

OPERATIONALIZING ETHICS IN LOCALIZED PRACTICES

In most cases, persons in charge of ethical aspects of design, development, and use—including notably also those in charge of compliance aspects—are uniquely positioned to operationalize fundamental values in their specific business or operational context. This section describes fruitful approaches to value implementation throughout the lifecycle of an AI system, focusing on more practical aspects of AI ethics. We will focus, in particular, on how values can and should be operationalized. In a nutshell, operationalization should initiate with the identification of the value base—including but not limited to the fundamental values discussed in this document—that should guide and constrain the design and use of AI systems throughout their lifecycle, and through which the relevant outcomes are to be evaluated. Operationalization then proceeds by defining, specifying, and implementing both technical and operational requirements that support establishing and maintaining conformity with those values throughout the lifecycle. We consider these tasks as best conducted—or at least managed—by the agents responsible for the ethical aspects of system design, development, or use.

> Persons or teams in charge of ethical aspects of AI system design, development, and use in a given organization should, together with internal and external stakeholders, establish a clear plan for operationalizing ethical values by first specifying a value base for the AI system and, second, devising a plan for implementing and protecting those values throughout system lifecycles.

Principles of AI ethics are not by themselves sufficient for addressing complex ethical issues. However, a systematically organized framework that includes AI ethics principles can "provide a valuable tool for detecting, conceptualizing, and devising solutions for ethical issues" [19]. To assist responsible agents in identifying and addressing ethical concerns, those principles must be carefully operationalized. The process of operationalizing AI ethics principles—sometimes also called "value alignment" [40]—consists roughly of two tasks. The first one involves identifying the core values that should regulate social behavior in a given context—including by extension the design and use of AI systems—and defining the value base of the technology in question accordingly. The second task consists of translating those values into practice. We will next provide information and more specific recommendations regarding these tasks. We will describe different aspects of operationalization as well as propose a practical approach based on existing proposals [7, 68]. We then discuss how the proposed framework, which is depicted on the following page, can be applied in ethical evaluations and analyses of trade-offs between different values.

PRO-ETHICAL AI PRACTICES:
A framework for operationalizing values throughout the AI system lifecycle

STEP 1. Define ethical values and principles.

- **Fundamental values:** What values should any possible AI system embody and reflect?
- **Laws and regulations:** What legal norms, regulations and standards apply to the AI system or the agent?
- **Context-specific values:** What values should be upheld in the specific use-context of the system?

STEP 2. Determine which values or principles should be prioritized over others (and when).

A) Weighted values: Should each value or principle be given a weight proportional to its importance?

B) Lexical ordering: Should the values and principles be ordered according to their priority?

STEP. 3. Establish and implement the following items for each value or principle.

Evaluation methods: How is the AI system's adherence to the value or principle estimated?

Quantitative instruments such as performance metrics and numeric indicators
Qualitative instruments such as case-studies, user interviews and different forms of feedback
Measures for absolute or non-comparative effects such as overall model error rate
Measures for relative or comparative effects such as model fairness

Targets: What is the normative target for a given value or principle?

A minimum acceptable value or a threshold that the system should always achieve
An ideal value that the system should achieved if possible

Specifications: What properties or functionalities should the AI system have to achieve the specified targets?

System specifications, quality requirements, safety requirements etc.
Requirements for applied data, models, interface, etc.

Operations: What processes and safeguards should be established to ensure the system achieves its targets?

Instructions, guidelines and safety protocols for human operators or end-users
Safeguards set up in the operational environment to constrain the AI system's behavior
Feedback mechanisms, procedures for users to seek remedy and redress

STEP 4. Use framework to monitor and remediate system behavior and impact throughout the lifecycle.

Review: Does the AI system continue to meet the targets established for it and the impact of its use?

Responses: What should be done to correct for violations and to ensure that the AI system continues to meet its targets?

Reflection: Do the established values and principles, in their operationalized form, capture all relevant requirements?

☑ Defining the value base of an AI system

The first task of value operationalization involves a process where a set of values and principles is transformed into a framework that informs, guides and structures action by specifying what ethical values and moral principles the AI system ought to conform to throughout its lifecycle (and which the people responsible for that lifecycle ought to follow). In other words, they compose the normative core to which technical and operational requirements—as practical operationalizations of those values established during the second task of "translation into practice"—correspond. The value base describes and prescribes, in other words, the constituents of "ethical success" in the context of a given AI system—or, more generally, a project or practice. They also function as the operative conceptual framework with which the negative and positive impacts of the product or service—in addition to any key activities that take place throughout the system pipeline—are assessed.

We have previously proposed a set of values and principles of ethics to be operationalized regardless of the use-context. This set included values such as the well-being of humans, non-human animals and the environment, human autonomy, freedom and dignity, justice and fairness. In the context of our framework, these are considered fundamental values—with corresponding, general principles of ethics—that can be accepted by different agents universally regardless of specific application context of AI. They are also considered important in light of the objective of protecting individuals' human rights and fundamental freedoms. However, we emphasize that the value base can and should also include values that reflect both general and context– or domain-specific requirements for compliance with legal norms and regulations. Good examples would include democratic ideals and social and collective values (such as participation, inclusion, diversity and solidarity) and other values recognized globally by various authorities (such as national laws and regional government organizations). Principles endorsed across professional communities, religions and cultures are also prominent candidates for inclusion in ethics frameworks. Depending on the use-context of the system in question, there may be also industry standards and professional codes of ethics, for example, which highlight values that AI systems should reflect in a given context of use. For example, medical and educational contexts may have certain specific values that responsible agents should uphold, which means that those values should also be included in the defined value base. Importantly, existing formulations of relevant values and principles that are found in legal documents or ethical codes, for example, can assist responsible agents in defining and operationalizing those values in the given use-context.

✓ Fundamental values: What values should any imaginable AI system embody and reflect?

✓ Laws and regulations: What legal principles should constrain the (behavior of the) system?

✓ Context-specific values: What values should be upheld in the specific use-context of the system?

☑ Translating values into practical norms, standards and criteria

The second task consists of translating abstract values into practice. Success in this task is crucially dependent on whether ethical objectives, requirements, and constraints have been specified in sufficient detail. After all, vague norms such as "one should do good" and "one should be fair" are difficult to translate into actionable procedures, clear guidelines, and system requirements and specifications to which different teams and individual persons working on (or with) the system can defer in their own work. One way to understand the aims of operationalization is to view it as a set of tasks that seeks to specify and establish action-guiding norms that different agents can follow in order to ensure morally acceptable (or even ideal) outcomes. Two types of norms can be distinguished in this respect [66]:

- "Ought-to-be norms". These norms describe what a system or a specific component, such as an algorithm or an interface, should be like, and what properties it ought to have to be acceptable or desirable from an ethical perspective.
- "Ought-to-do norms". These norms describe what human agents, such as developers and system operators, ought to do in order to ensure that the system or the outcomes of its use are acceptable or desirable from an ethical perspective.

Ideally, both types of norms should be specified in a way that covers the entire lifecycle of the AI system. For example, at each stage of the lifecycle, responsible agents—be it designers and model builders or persons operating or maintaining the system—should know what is required from them in light the established ethical objectives and constraints.

Importantly, moral values cannot be observed "directly", which is why operationalization should also lead to a concrete set of standards and methods that can be used to evaluate AI systems from an ethical perspective. For example, practitioners should have access to metrics and indicators that can be used to generate information about the ethically relevant aspects of the technology. One approach is to use so-called Key Ethics Indicators (or KEIs), which are similar to key performance indicators (KPIs) but "provide a more holistic understanding of whether or not an algorithm is aligned to the decision-maker's ethical values" [68]. They can complement standard measures of performance and success that might be applied at an organizational level to evaluate its performance. KEIs can be tracked with different quantitative and qualitative methods (such as statistical model performance measures or regularly conducted surveys and interviews). However, there should also be a way to observe whether and when a given value or principle—or more specifically, the KEI that is used to track it—has in fact been violated. One way of going about this is to attach a criterion or a threshold to each KEI, which establishes a "no-go" condition for the KEI in question. Violation of a criterion or a threshold should trigger an investigation into the cause of the violation or failure, and lead to actions that mitigate the relevant harm. For example, violation of a KEI might lead the system to power down or, alternatively, initiate a "fallback plan" where a duplicate system or a human operator takes control of the process.

A systematic and well-defined framework for ethics can help development teams and responsible practitioners identify ethical issues, risks and tensions arising in the context of their specific work tasks and problem areas. With the help of a clear framework, responsible agents can better identify what values, rights or freedoms are at stake when a problem arises, for example. An ethics framework can also be used to broaden standard notions of performance and success in business and technology development contexts by way of bringing diverse forms of value into the picture. By using a framework that includes various constituents of "ethical success", responsible agents can conduct more holistic evaluations of their products, services and systems, as well as communicate those constituents better to relevant stakeholders (both internal and external) throughout the AI system's lifecycle.

Noting that principled approaches to AI ethics have been criticized for being useless [83] and non-action-guiding [80], responsible agents will have to make sure that their framework can be of assistance in identifying and responding to ethical issues in practical contexts of technology development and use. In particular, we suggest that there are two goals that responsible agents should keep in mind when operationalizing values and principles of ethics for AI system lifecycles in their specific context of development and/or use:

- The epistemic function of the ethics framework: Responsible agents should be able to use the ethics framework as a "diagnostic tool" to identify whether and when specific actions or outcomes—such as an AI system's behaviors, model properties, or consequences of system use—are permissible, impermissible or obligatory.

- The pragmatic function of the ethics framework: Responsible agents should be able to use the ethics framework—or individual principles included therein—as a decision-making procedure [87] when deciding how to realize certain values, how to address conflicts between values or principles, and how to respond to ethical misconduct.

The overarching aim, in other words, is to operationalize the ethics framework in a way that ensures it can generate consistent, coherent, and determinate (a) ethical evaluations and (b) prescriptions for action. We propose that four sets of items (see below) are defined in connection to each value or principle included in the framework to achieve this aim.

Targets: What would the behavior and impact of an AI system ideally be like?

Specifications: What properties should the system have to reach the normative targets?

Operations: What safeguards and procedures are needed to achieve the normative targets?

Evaluation methods: What measures and methods can be used to assess success in meeting targets?

Targets. For each value or principle included in the ethics framework, responsible agents should define a target that should be achieved. Furthermore, the relative weight of each target should be specified, or they should be ordered according to the importance of each value or principle (and target).

The identified values and principles comprise the "value base" of the system lifecycle—they are the values and norms an AI system should embody and which it should be used to promote and uphold. To be action-guiding, however, those values need to be translated into well-defined objectives and targets. Responsible agents should thereby define and specify normative targets which determine a normatively preferred outcome. In practice, targets may include, for example, minimum acceptable rates of true positive predictions or maximum percentages of negative user feedback. They may also include targets for qualitatively estimated impacts, such as the system's estimated effects on user autonomy or freedom. As targets are essentially proxies for specific values included in the value base, they have to be defined with respect to each value individually. Defining these targets, responsible agents should ask:

- What is the best possible outcome that can result from the use of an AI system?
- What should the behavior of an AI system ideally look like?
- What is necessary for a given value to be realized in the use of the system? What is sufficient?

Formally speaking, there can be different types of targets. "Perfectionist" targets will specify "ideal" or "perfect" outcomes—the best possible outcome in terms of fairness or harm-avoidance, for instance. While perfectionist targets are difficult to achieve in practice, specifying such targets can help agents identify deviations from the envisioned best -case scenario. Targets can also include sufficiency conditions that define an "acceptable minimum" or "acceptable maximum" to be achieved with respect to a given value, such as a maximum amount of prediction errors. These types of targets are most applicable when there are clear, well-defined standards for what counts as a violation of some ethical value or legal norm.

Determining which values or targets constitute priorities. To resolve value conflicts and trade-offs that may arise, the relative importance of each value and principle—and, by extension, their attached targets—should be determined. Depending on the use-context, values and principles might have to be given "weights" that specify their relative importance, for example, or they might have to be ordered in terms of their significance. While determining weights or the order of values is difficult and demands care, it is necessary in order to ensure the ethical framework is action-guiding in practical situations. For example, one might expect that the predictive accuracy and fairness of a given algorithm will involve a trade-off, and thereby one should establish—with due regard for context and legal norms—whether in cases of conflict one should prioritize the accuracy or fairness of a machine learning model.

Specifications. For each target, responsible agents should devise a list of system specifications and technical requirements that are necessary for the system to achieve that target.

Specifications concern the AI system specifically as a technical artefact in that they include prescriptions regarding the properties of the applied datasets, algorithms, code, user interfaces, and other technical components or aspects of the system. In other words, specifications describe what should the system be like in order to realize the values included in the value base, and to achieve the defined targets. Recalling the notion of "ought-to-be norms" discussed previously, specifications defined in relation to ethical values and moral principles should describe what the system ought to be like in order to fair, beneficial, safe, and so on. Defining specifications, responsible agents should ask:

- What properties should the AI system or a component have from an ethical perspective?
- How should the system function in order to realize important values and achieve the established targets?
- What technical properties would contribute to valuable outcomes in the use of the system?

Operations. For each target, responsible agents should define a set of procedural guidelines and safeguards that serve to ensure that the system meets its targets.

Operations refer to the procedures, guidelines, mechanisms and safeguards beyond the technical system itself that should be in place to ensure that the system meets its targets. Operational requirements prescribe what humans interacting with the system are obligated and permitted to do. In this sense, operational requirements are more akin to "ought-to-do norms": they guide and constrain the actions of humans that are responsible for building or using the system, for example. They may also concern what people subjected to (or affected by) the AI system should know and do when dealing with the system, such as how the system ought to be powered down safely or how the outputs of the system ought to be reviewed in order to ensure they are correct. Defining these operational requirements and safeguards, responsible agents should ask:

- What should persons building, using or interacting with the AI system know or do?
- What safeguards should be implemented for specific procedures during system development, testing or use?
- What objectives and constraints should be explicated to ensure that developers know what is expected of them?

Evaluation methods. For each target, responsible agents should establish a suitable method of evaluation that can be applied to track whether the system achieves that target.

To enable continuous monitoring of the AI system—including also the processes that go into developing, building, operating and maintaining it—responsible agents should specify what methods and measures are used to evaluate whether and to what extent the targets are met. This calls for evaluation methods and means of observation (and associated criteria) that can be used to generate information about the system's ethical aspects. Various metrics can be used for generating information about expected or actual impact. For example, quantitative measures and metrics can be best applied to estimate things such as model performance, but qualitative instruments ranging from surveys and interviews to more narrow instruments such as user feedback can be used to assess things such as how the AI system affects users' perceptions of their own autonomy.

Importantly, evaluators also need standards against which the identified types of impact are assessed. Evaluative standards should reflect the established targets. Defining evaluative methods and standards for an AI system or a specific use-context, agents should ask:

- What evaluation methods should be implemented to enable specific persons to evaluate the system?
- What metrics and observational methods can be applied to identify positive and negative effects?
- How could the system's impact be monitored continuously in terms of the fundamental values?

Next we proceed to discuss questions related to ethical evaluation and review in the context of AI systems.

ETHICAL EVALUATION OF AI SYSTEMS

In previous sections, we have noted that ethical practices need to be explicit and standardized in order to be effective. The proposed model for operationalizing values and principles of AI ethics can and should be used to both track and evaluate morally and legally relevant aspects of AI systems throughout their lifecycle. As part of the central activities delegated to persons or teams responsible for ethical aspects of design, development and use, we suggest that the ethics framework defined and implemented by responsible agents should be used to conduct three different types of evaluations of AI systems and their effects (throughout development and use-phases). These different evaluations are described briefly below. They can also be used to complement impact assessment procedures—for example, responsible agents conducting an assessment of the equality impact of deploying a given AI system could complement the overall assessment with a targeted evaluation of the applied model's fairness or with a trade-off analysis. On the following pages, we also provide some theoretical resources, recommendations and a list of prompts that can help responsible agents conduct thorough ethical assessments.

1 COMPREHENSIVE EVALUATION

A comprehensive evaluation consists of an overall evaluation of the AI system, covering all values included in the ethics framework or established "value base". The AI system is evaluated against all targets using the determined methods of evaluation. The evaluation should generate an overall assessment of the morally (and legally) relevant aspects or impacts of the system in question.

2 TARGETED EVALUATION

Targeted evaluation refers to an evaluation of the AI system from the perspective of a single value or principle included in the ethics framework or "value base". The AI system is evaluated against a set of targets that are attached to a given value or principle using the determined methods of evaluation. The evaluation should generate an assessment of the AI system or its impact in light of the chosen value or principle.

3 TRADE-OFF ANALYSIS

Trade-off analysis refers to an analysis which aims at identifying value tensions or value—in particular, whether increasing or decreasing one value (or one target) increases or decreases another value (or another target). By looking at pairs of values and their corresponding metrics or indicators, trade-off analyses can reveal whether the AI system has mutually exclusive desirable properties or types of impact which are in tension with one another.

A PERSPECTIVE:
Identifying and addressing trade-offs

Trade-offs refer to cases where an agent has two or more values which they should realize simultaneously, or two or more obligations that they should fulfill, but realizing those values or fulfilling those obligations is not possible for some reason. Specifically, the term "trade-off" has traditionally covered cases where increasing a value X leads to a decrease in another value Y. When faced with a trade-off, an agent has to trade one value for another, in other words. Designers, developers and users of AI systems may encounter a variety of trade-offs that can present agents with hard choices. For example, a common trade-off that occurs when training a machine learning model is the so-called bias-variance trade-off, which refers to the well-known problem that a model cannot usually both capture the regularities of the training data and simultaneously generalize well to unseen data. Research on AI ethics has also identified various types of trade-offs that can arise when training, selecting or refining machine learning models. For example, it has been demonstrated that it is impossible to train a machine learning model that satisfies different definitions for what makes a model fair [22, 62]. Similarly, there may be trade-offs between model security and model transparency, because an explainable machine learning model can be more vulnerable against model attacks, where an attacker seeks to infer sensitive information by probing the model.

Trade-offs and tensions between values are a rule rather than an exception. Responsible agents should be prepared to encounter and address them. First, persons in charge of ethical aspects of AI system design, development or use should seek to actively identify trade-offs. This implies that there should be processes or practices (such as ethical review and impact assessment) that are focused in part on identifying trade-offs between values in the context of AI system development and use. Once a trade-off is identified, responsible agents should also address them properly. They should establish what kind of a trade-off they are dealing with because the nature of the trade-off can in part determine whether and how that trade-off can and should be resolved (see below). When the nature and scope of the problem is clear, the trade-off in question should be addressed with due regard for context and applicable legal norms and regulations, and in an effective and feasible manner. Once the trade-off is addressed, responsible agents should make sure that the identified trade-off and the measures taken to address it have been documented in detail.

It might be useful to note that there are at least three types of trade-offs that can arise between different values or obligations: First, an inherent trade-off arises when there are two or more conflicting values, duties or obligations which cannot be realized or pursued simultaneously, or when pursuing an increase in one value comes always and necessarily with a cost to another. In other words, the values or normative principles are not logically compatible even in theory—one cannot have both. If an identified trade-off is an inherent trade-off or a logical trade-off, the agent responsible for addressing the trade-off typically has to choose which value or obligation to prioritize. Second, a practical trade-off does not involve values, duties or obligations that are inherently incompatible in principle. Rather, when a practical trade-off arises when realizing those values simultaneously or fulfilling the relevant obligations is highly unlikely, significantly costly or otherwise infeasible due to the current contingent, practical circumstances. There is a "background reason" or "background cause" that makes two values or obligations mutually incompatible. Practical trade-offs include rather common instances of moral dilemmas, where two valuable things cannot be achieved simultaneously (even though they are theoretically compatible). If an identified trade-off is practical in this sense, the agent responsible for addressing it should consider whether the trade-off could in fact be addressed by removing the "background reason" or "background cause" that gives rise to trade-off. For example, some theorists have noted that trade-offs between definitions of model fairness are not theoretically incompatible but rather present us with a practical trade-off [15, 37]. Lastly, a false dilemma arises in cases where an agent perceives there to be a trade-off between values, duties or obligations but the agent in fact has available a possible third option or course of action which could be chosen. The perceived tension dissolves when the third option is identified. False dilemmas are not trade-offs at all, in other words, but rather involve a misunderstanding or a misperception of a given issue. When an agent identifies a trade-off, they should first carefully consider whether they are in fact dealing with a false dilemma. Sometimes reconceptualizing or reframing a problem can prove useful in this respect—looking at things from another perspective may show us that there is a problem where there previously was not, and vice versa.

When we discussed the lifecycle ethics approach, we noted that ethical evaluations should not be restricted to narrow frames, such as the evaluations of model performance or the quality of the data, even though it is necessary to evaluate also the aforementioned things. This is because narrow evaluations—even if necessary to generate a holistic picture of the ethical aspects of AI systems—can remain myopic in the sense that they neglect larger socio-technical and political contexts, for instance, which should bear on the responsible agents' judgment about moral valence. For example, researchers have argued that ethical evaluation in the context of AI systems should account for the various features and complex dynamics of sociotechnical systems, such as their tendency to adapt and change over time, as well as different temporal perspectives, such as long-term effects of using AI systems [37, 70, 98].

Responsible agents should employ diverse ethical perspectives and sociotechnical lenses—including by way of combining technical analysis and social systems analysis, for example—to generate comprehensive and truthful evaluations of AI systems' ethical aspects.

The practical necessity of employing diverse ethical perspectives and sociotechnical lenses during ethical evaluations is grounded in the fact that an assessment of an AI system—or its technical properties or components—cannot generate a holistic and truthful picture of the ethical dimensions related to its use. As an illustrative example, a facial recognition algorithm might be nominally fair in the sense that its expected rate of mispredictions is roughly equal between two demographic groups A and B. However, if a third party who has access to the algorithm only uses it on individuals who belong to group B—with the intent of subjecting members of B to undue scrutiny, for instance—the results of the actual use of the algorithm are nonetheless comparatively unfair. An assessment of fairness as it relates to the facial recognition system, in other words, is not exhausted by an evaluation of model fairness (i.e., misprediction rates) because the unfair effects arise in the actual use of the system. Furthermore, precisely because the effects arise in such a manner, responsible agents would have to address the issue by imposing restrictions on "downstream use" of the facial recognition system, for example. The lesson here the larger sociotechnical context in which an AI system is embedded and used—including the general and unique features of that context—has to be taken into account.

This key lesson has been also demonstrated its importance in the context of AI safety. Many risks, negative effects or beneficial types of impact cannot be identified without looking at things from a broader "systemic" level. For example, many of the system-level risks and effects—such as accidents resulting from the use of AI systems or other events that are relevant from a legal or moral perspective—cannot be anticipated by merely looking at system components and their reliability, for instance [30]. Many types of risks and relevant causes of errors can be best identified and managed by introducing redundancy at different levels—such as by implementing both technical safeguards and operational safeguards—and by modeling interactions between the AI system, its users and operators, and the sociotechnical use-environment, for example. On the following page, we have listed some features and dynamics of sociotechnical systems alongside both examples that illustrate their relevance for ethical evaluation and intervention design and practical lessons that responsible agents should follow.

PRO-ETHICAL AI PRACTICES:
Ethical evaluation and system design under sociotechnical complexity

FEATURES OF SOCIOTECHNICAL SYSTEMS	EXAMPLES	TAKE-AWAY
Complexity Sociotechnical systems involve numerous interacting agents, technologies, environmental factors, and social structures.	AI systems have (in)direct effects on agents and processes in a given use-context. These effects are mediated by multiple factors—often latent ones that are not found in applied datasets.	Responsible agents should consider and model the complexity and dynamics of the AI system's sociotechnical use-context, including latent factors and pathways of influence between (non-)agents.
Distributed information and control Access, power, control, and information are distributed across different (non-)agents (such as persons and technologies) in the sociotechnical system.	Stakeholders (such as developers, end-users, decision subjects) may require different types of explanations for an AI system's outputs due to differences in informational interests and expertise.	Responsible agents should identify and model the informational and operational requirements, affordances and limitations of different (non-)agents.
Levels of organization Agents and behaviors in the system are organized in different ways. Interventions on one level can have effects that aggregate upwards, trickle downwards, or both.	A recommender system which structures consumption behavior among users can create—through aggregation and cumulation—negative effects on service providers or product developers.	Responsible agents should consider and model how vertically layered and interdependent behaviors and processes within the sociotechnical system can mediate or affect the behavior of the AI system.
Stochasticity and non-linearity Random events within a sociotechnical system introduce unpredictability. Interactions between (non-)agents within the system may not be linear, regular or continuous, for example	Weather conditions, such as storms, may affect or transform the environment in which an embedded AI system—such as an autonomous vehicle or a delivery robot—operates, and thereby create new or unexpected safety hazards.	Responsible agents should take into account that latent factors and sources of randomness may affect how (non-)agents behave in the actual use-context of the AI system.
Emergence Interactions between (non-)agents and features of the sociotechnical system can lead to surprising or unpredictable outcomes.	When Microsoft deployed the Twitter chatbot "Tay", Twitter users taught it to post racist and antisemitic tweets. The undesirable consequences emerged from the interaction between the system and the platform's users.	Responsible agents should note that the deploying and using an AI system can create negative externalities and unexpected consequences, when the system interacts with (malicious) agents or the use-environment.
History (Non-)agents within the system are the product of past interactions and store information about the past.	Data pipelines contain sources of noise, error, and bias. Taking information about the past as "ground truth", an AI system can end up reifying errors and embedding societal bias into algorithmic decision-making policies	Responsible agents should account for the causal history of applied datasets and other components, and research the institutional and/or historical context in which the AI system is used (or planned to be used).
Adaptivity and evolution (Non-)agents and behaviors adapt to changes in the system, possibly changing how the system as a whole operates in the future.	People can wear face masks to prevent identification by a newly introduced facial recognition system. Contexts and environments also change, introducing data and concept drift, for example.	Responsible agents should anticipate and/or model whether and how affected (non-)agents and processes will (or could) adapt to the use of the AI system.
Time-delays Interventions on system (non-)agents, processes, or structures may create effects that do not manifest immediately. Expected and unexpected effects may emerge slowly through time or with delay.	Using an algorithm that has been tested for fairness can, after multiple rounds of use, generate long-term impacts that are no longer fair according to the same standard [70].	Responsible agents should anticipate and identify negative and positive long-term effects of AI system use via modeling or simulating them, for example.
Feedback-loops Information created with the AI system can create effects that feed back into the system later on, thus creating looping effects and self-reinforcing cycles.	Use of biased data in predictive policing can focus police presence on geographical areas inhabited primarily by racialized and marginalized communities, leading again to more arrests in those areas [33].	Responsible agents should anticipate and identify looping effects, and seek to ensure that AI systems do not create self-fulfilling cycles that have negative effects or which are counterproductive to social values and aims.

The previously described features of sociotechnical systems show how an overly narrow mindset that is focused on technical components and simple metrics, for instance, can result in uninformed evaluations or ineffective designs for safeguards. Recognizing and accounting for a broader set of social, technological and user-related factors, including many other and often latent effects and sources of stochasticity, is a crucial part of what it means to be sensitive to the sociotechnical complexity that pertains to AI systems and their safe and lawful use. To demonstrate the relevance of the previously listed factors, we can consider autonomous cars as an illustrative example of how the different features of sociotechnical systems are relevant from the perspective of AI system safety.

TABLE. Accounting for the features and dynamics of a sociotechnical use-context: An example from autonomous cars

Complexity. The behavior of the autonomous car is affected by the actions of the user who exercises control over the vehicle, pedestrians whose actions are tracked by the vehicle's sensors, as well as environmental factors such as the weather which affects things such as the condition of the road and visibility.

Distributed information and control. The behavior of the autonomous car is affected by the input it receives from its external and internal sensors such as from camera image and from the user. Pedestrians also interpret the vehicle's actions to move safely in the streets. Control over the flow of traffic is distributed among drivers, autonomous vehicles, pedestrians and traffic lights, for example.

Levels of organization. The overall functioning of traffic on the road is the joint product of the behavior of different agents, technologies and environmental factors. The behavior of one driver can affect other drivers' actions, such as when they change lanes. Overall traffic also constrains each driver's actions separately, slowing down their driving, for example.

Stochasticity and non-linearity. Unexpected events—including things such as jaywalking pedestrians, animals walking on the road, or malfunctioning traffic lights—can affect the autonomous car's behavior and create risks for safety. In addition, random events, such as the vehicle's driver spilling a drink on the dashboard, may also create abrupt risks.

Emergence. During a longer time period, human passengers who observe that the autonomous vehicle is driving consistently and safely might incrementally become over-reliant on the vehicle's capabilities and thereby pay less attention to the road.

History. The dataset used to train the autonomous vehicle might be more representative of urban environments than rural environments, or more representative of right-hand traffic when compared to left-hand traffic. Consequently, the autonomous vehicle might perform better in urban environments or in geographical locations where traffic is on the right-hand side.

Adaptivity and evolution. Both drivers of autonomous vehicles and those driving on roads with many autonomous vehicles might develop novel behaviors that protect them from safety risks that are characteristically created by autonomous vehicles. Malicious agents who are up to no good might start to trap autonomous vehicles into parking lots by painting white lines onto the ground.

Time-delays and feedback-loops. Drivers of traditional non-autonomous vehicles might observe autonomous vehicles' behavior over time and learn to better anticipate their behavior. If autonomous vehicles' learn online during use, they might also pick up the adaptive patterns that develop amongst drivers of non-autonomous vehicles through time.

6 PRACTICING “ALGORITHMIC ACCOUNTABILITY”

What practices should be adopted?

AN OVERVIEW

Given the increasing impact that AI systems can have on individuals' rights and freedoms, political institutions, democratic processes, and society at large, it is clear that the development and use of AI systems should adhere to applicable laws and regulations but also meet publicly accepted and shared ethical and social standards. These standards are often understood as standards for so-called "algorithmic accountability" which refers to accountability relationships where multiple actors—including decision makers, developers, system operators and end-users, for example—are obligated to explain and justify their design or use of their AI system, including decisions that have affected its properties or outcomes of its use [119]. "Algorithmic accountability" should be understood as analogous to political accountability, a concept often discussed in relation to the responsibilities of political institutions and their representatives in relation to affected citizens: "[just] as democratic citizens have the right to scrutinise and hold to account the exercise of political power, so algorithmic constituents have the right to scrutinise and hold account the exercise of algorithmic power" [14].

Accountability is pursued, enforced, and demonstrated through what we call accountability mechanisms. A recent report by the Ada Lovelace Institute has reviewed existing mechanisms that have been proposed in the research literature on AI ethics and governance, for instance [1]. These include, for example, codes of ethics for AI practitioners, oversight and auditing, impact assessment processes, various mechanisms for ensuring affected individuals' access to redress and remedy as well as the establishment of certification and registration systems for automated decision-making systems. Importantly, designing effective accountability mechanisms for the context of AI systems is notably difficult due to various factors including, for example, the number of agents and stakeholders involved, the modular, adaptive and complex nature of AI systems, and the messiness of software development pipelines [119]. Effective mechanisms require well-defined institutional incentives and enforceable legal frameworks, with additional incentives and an external pressure for organizations' to demonstrate accountability being ideally introduced by things such as media coverage and civil society activism [1].

In this section, we will dig deeper into "best practices" and (self-)regulatory mechanisms for promoting algorithmic accountability. Ideally, these practices would be mandated by law, but we argue that responsible agents should implement them regardless of whether they are (presently or in the future). This section shifts focus into more practical aspects of AI ethics, complementing our previous recommendations on laws and regulations, organizational practices, and operationalization of ethics. Our discussion will focus mostly on impact assessment and review mechanisms, although we will discuss things such as documentation and responsible communication as well. There are many other existing proposals which would be worth discussing here, but alas we will have to focus on those considered most important for present purposes. An overview of our proposal is depicted on the following page.

TABLE. Examples of best practices for algorithmic accountability in the AI system lifecycle

PHASE OF AI SYSTEM LIFECYCLE	EXAMPLES OF BEST PRACTICES AND ACCOUNTABILITY MECHANISMS
CONCEPTUALIZATION AND DESIGN Conceptualization and design of the AI system Problem formulation and stakeholder identification System specifications and operational requirements	**Ex ante review of the use-case, concept or design** Assessment of acceptability, necessity and proportionality. **Operationalization of ethics framework (see section 5)** Definition of values and principles and their corresponding targets, evaluation methods, technical specifications, and operations and safeguards.
DATA COLLECTION AND PRE-PROCESSING Data capture and collection Data pre-processing and labeling Data storage and transfer	**Dataset evaluation** E.g., data suitability and applicability, accuracy and integrity, representativeness, bias and harmful content, and other contextually relevant aspects
MODEL BUILDING AND EVALUATION Model training and hyperparameter tuning Model evaluation and validation System testing and prototyping (in vitro)	**Model evaluation** E.g., model performance and accuracy, model security and vulnerabilities, model fairness, energy consumption and emissions
USE-PHASES OF THE SYSTEM System testing and prototyping (in vivo) System deployment and use System maintenance and termination of use	**Impact monitoring** E.g., comprehensive and targeted ethical evaluations, trade-off identification and analysis, and algorithm audits (regularly or upon request)
	Impact assessment: E.g., HRIA for high-stakes decision-making contexts, and including computing-related impacts to DPIA, EQIA and EIA. **Ethics-by-design:** E.g., implementing transparency requirements and risk management procedures **Documentation:** E.g., registering AI systems, documenting applied datasets, algorithms, and models, and documenting computing-related impact for impact assessment and auditing purposes.

TABLE. Leveraging existing impact assessment frameworks for assessing algorithmic impact

A GENERAL FRAMEWORK FOR ALGORITHMIC IMPACT ASSESSMENT	
1. HUMAN RIGHTS / FUNDAMENTAL RIGHTS IMPACT ASSESSMENT Necessary for AI systems used in high-stakes contexts Voluntary for AI systems used in low-stakes contexts Existing frameworks determine the structure and deliverables of the impact assessment process. **2. DATA PROTECTION IMPACT ASSESSMENT** **3. EQUALITY IMPACT ASSESSMENT** **4. ENVIRONMENTAL IMPACT ASSESSMENT** Applicable laws, regulations, and standards determine necessity, structure, and deliverables.	**ALGORITHM AUDIT (1-4)** A technical audit of the AI system should be included into each impact assessment process to estimate impact resulting from computing and/or information processing. The methodology of the technical audit will depend on the use-context and nature of the AI system in question. Other empirical methods, including qualitative methods, should estimate operational impact as mediated by AI-user interaction, when possible.
Impact assessment required ex ante and regularly throughout use-phases of the AI system. Includes traditional stages of impact assessment, scoping, a baseline study, consultation of affected stakeholders, etc. Impact statements (or reports) include a separate section for computing-related impact.	

EX ANTE REVIEW AND JUSTIFICATION

A key lesson learned from past failures with AI system deployment as well as existing cases of ethically or legally problematic data science projects is that many of them lacked a well-defined, scientifically and socially supported, and context-appropriate objective or purpose of use. They went wrong from the start, or at least quite quickly. For example, a recent report documenting the failures of automated decision-making systems used in public contexts shows that about a half of existing systems that were applied and scrapped were in fact scrapped after—and not prior to—deployment [92]. The report suggested that primary reasons for why the use of automated decision-making systems was ceased include, for example, those systems' lack of effectiveness, public criticism and contestation, as well as critical media investigations that exposed their problems. What many existing cases of failures with regard to AI system deployment show is that neglecting ethical and legal considerations related to the acceptability, necessity and proportionality of developed or deployed AI systems can be highly risky or even dangerous, but also immensely costly.

Establishing well-defined, contextually grounded, and scientifically supported use-case for an AI product or service is of utmost importance from a legal and ethical perspective. This idea is also widely supported by research on AI ethics and accountability. Part of the support stems from the crucial observation that some applications of machine learning—such as "snake oil AI" applications which were discussed in relation to the principle of nonmaleficence above—simply do not and cannot work, and that they fail to do so precisely because they are designed to solve tasks that are scientifically dubious or even conceptually impossible [85]. Yet due to misplaced trust on technology and misconceptions about the "neutrality" and "objectivity" of AI systems, for example, such applications may nonetheless enjoy a level of mistaken trust-by-default. This problematic dynamic is also arguably fueled by technology developers who are incentivized to downplay the shortcomings and limitations of their technologies in order to maximize profit, for example, as well as unfounded and excessive "AI hype" sustained by both industry agents and the media. In many cases, the problems that machine learning applications are envisioned to solve are also so complex and multifaceted that they cannot be addressed only with technological means—perhaps at all. Consequently, the so-called "functionality assumption"—namely, the prevalent assumption that AI systems can actually do what they are expected or intended to do [91]—should not be the unquestioned premise of AI system development and use. Rather, whether a given AI system constitutes both an appropriate and effective solution to a given well-defined problem should be in all cases assessed carefully without prejudice and bias. The burden of proof, we suggest, should be on the party who is proposing a technological solution to a socially or politically complex problem, or any high-stakes problem to which automation is envisioned to provide an answer. Unless it can be clearly demonstrated that there is no acceptable and clearly better alternative than to use an AI system, such as a more effective non-technological solution, no "license to operate" should be granted—at least without further justification.

Drawing lessons from past failures pertaining to the deployment automated decision-making systems, the authors of the previously mentioned report recommend the implementation of impact assessment procedures, legality reviews, and shifting the burden of proof with regard to justifying the development and deployment of automated systems onto the party wishing to deploy the system [92]. They also emphasize the importance of understanding specific contexts and areas of life where the use of automated decision-making is considered unacceptable by the public, and urge developers and public organizations to consult and investigate whether and how automated systems should be used in the first place.

Recommendations on ex ante review and justification

Taking into account the previously discussed suggestions, we recommend adopting the practice of subjecting AI system use-cases (and data science projects) to ex ante legal and ethical review. An ex ante review requires the responsible agent—e.g., developer, operator or other relevant party—to provide evidence for the (expected) effects of the envisioned AI system prior to full-scale deployment. In particular, responsible agents should determine and provide evidence for whether the defined use-case for an AI system or automated decision-making practice, for instance, can be justified in light of commonly accepted and well-established public standards as well as context-specific requirements for algorithmic accountability (including requirements necessary for a "social license to operate"). Specifically, the requirement for ex ante review implies a process for reviewing the use-case against a set of legal and ethical criteria related to (1) the system's envisioned objectives and goals and (2) the necessity and proportionality of the system as a means to achieving those objectives and goals.

A requirement to review an envisioned AI system use-case—or data science project, more generally—constitutes a feasible and desirable mechanism for both assessing and demonstrating algorithmic accountability. It seeks to ensure that public reason and standards of accountability which ought to be upheld in democratic societies function as effective constraints on technology development and use [14]. The need for ex ante review is arguably most pressing in cases where an AI system is envisioned to operate in a high-stakes context, such as a public or (semi-)public decision-making context. A requirement for ex ante review in such cases could (and should) be implemented by policymakers and regulators at the level of policy and regulation. However, responsible technology developers and operators could also implement ex ante review processes for self-regulatory purposes—for example, as a part of quality and compliance management processes. An ex ante review would ideally take place at the initial stages of the AI system lifecycle, such as when there is a description and/or a proof-of-concept regarding the system or an AI-based product or service. However, we emphasize that the review can require agents to address questions related to technical features of the AI system as well, meaning there can be a prototype or demo ready, for example. The review would establish whether moving forward with development is warranted. A positive or accepted review of the envisioned use-case would provide a warrant to move forward whereas, in case the product or service concept is demonstrably infeasible or impermissible for legal or moral reasons, the review process would foreclose that option without the responsible agents making necessary revisions or implementing a risk management strategy, for example.

RECOMMENDATIONS:
Ex ante review and public justification

AI system use-cases, concepts and designs should be subjected to ex ante review

Our recommendation is that policymakers should implement a requirement mandating that any AI system envisioned to operate within a high-stakes decision-making context should, during or after the initial phase of product/service conceptualization or design, undergo an ex ante review of the AI system use-case (including a concept or design for an AI-based product or service or a data science project). The review should evaluate and demonstrate whether the envisioned system has an independently acceptable, well-defined and tractable, scientifically sound objective or goal, and whether using the system to meet that objective or goal is necessary, proportional and preferable when compared to alternative means. Where applicable and necessary, ex ante review can also be conducted as a part of a broader impact assessment procedure. If ex ante review is not required by law, it is recommended that responsible agents conduct one in case they are developing an AI system to be used in a high-stakes decision-making context. An ex ante review of AI system use-cases, AI-based products/services, or data science projects should initiate with the organization developing, operating or controlling the system determining whether the defined use-case (or AI system) meets criteria of ex ante justifiability such as acceptability, necessity and proportionality. Notably, the process may require the responsible agent to review any initial system specifications that have been established or technical prototypes which might have been built, for example. Respectively, we emphasize that even though the review is to be conducted before deployment, it may involve an assessment of the technical aspects of the developed or to-be-developed AI system. Some prompts to structure the review are provided below.

1. REVIEWING THE OBJECTIVE OR GOAL OF THE AI SYSTEM

- ☑ Is the objective/goal acceptable? Why?
  E.g., is there an independent moral and legal justification for pursuing or achieving the objective?
- ☑ Is the objective/goal well-defined and computationally tractable? Why?
  E.g., does the system execute a clearly defined and simple part of a larger task?
- ☑ Is the objective/goal conceptually sound and scientifically valid? Why?
  E.g., is there a successful previous use-case or scientific evidence in support of using the AI system for the defined task?

2. ASSESSING THE NECESSITY, EXPECTED EFFECTIVENESS AND PROPORTIONALITY OF THE AI SYSTEM

- ☑ Can the defined objective/goal be achieved by using the system? How?
  E.g., is it possible to use the system robustly and reliably as a means of achieving the desired outcomes?
- ☑ Can the system's expected effects be clearly delineated? What are they?
  E.g., does the system generate a specific set of desired outcomes even under uncertainty and changes in its use-environment?
- ☑ Can the system's expected effects be achieved without creating disproportionally harmful effects? How?
  E.g., can the system be used without compromising other values, violating rights or generating unjustifiable burdens?
- ☑ Is using the AI system preferable to alternative means of achieving the defined objective/goal? Why?
  E.g., what benefits does using the AI system have when compared to other methods, including non-technological ones?

FRAMEWORK FOR IMPACT ASSESSMENT AND AUDITING

A key requirement from the perspective of algorithmic accountability relates to the identification and evaluation of effects resulting from AI system use and key choices and decisions regarding the design of the system in question. However, many traditional quality and risk management processes (e.g., evaluations of component reliability in the context of software development processes) are not sufficient to ensure that individuals' rights, freedoms and well-being are not compromised as a result of AI system use. The use of increasingly more effective and ubiquitous data processing techniques, for example, forces data controllers to "go beyond the traditional focus on data quality and security, and consider the impact of data processing on fundamental rights and collective social and ethical values" [74]. As many AI systems are updated over time and even learn continuously from users' behavior, moreover, assessments of risk and impact that are conducted at a given point in time are not sufficient to ensure that novel types of negative outcomes do not result from system use down the line. This has been a key insight driving research on methods for monitoring and mitigating unacceptable biases that AI systems learn through time [70]. For these and other reasons, the idea that technology developers and/or organizations that use AI systems with considerable social, legal or economic consequences should be required to assess, regularly monitor, document, and disclose "algorithmic impact" has gained traction [78].

Given the myriad of ways in which AI systems can affect individuals and communities, proposals for impact assessment frameworks have varied in the envisioned scope of the assessment. Existing proposals have included both frameworks for assessing many different kinds of effects—including the overall impact of an AI system—as well as for estimating impacts of a specific kind, such as a given AI system's impact on equality [55]. Impact assessment processes are often envisioned to include a technical review of the AI system under assessment—or so-called "algorithm audits" or "algorithmic audits" [29]. Algorithm audits could focus on exposing unacceptable, systematic biases in machine learning algorithms or trained models, for instance, generating also recommendations on how to mitigate such biases, among other issues.

In this section, we will propose and discuss our general framework for impact assessment and algorithm auditing as means of protecting rights and promoting algorithmic accountability. We discuss the design and implementation of the framework. Specifically, we provide recommendations on four different kinds of impact assessment processes focused on different types of impact, discuss how algorithm audits could be included in those processes. These discussions are meant to provide actionable guidance for practitioners and researchers by pointing out applicable evaluation methods and metrics, for example, and by identifying actions that could be taken throughout the AI system's lifecycle to mitigate salient types of negative effects.

Impact assessment

Impact assessment refers to a procedure for identifying and responding to potential or actual impacts of an existing or proposed AI system (or other decision-making system). Impact assessment procedures commonly also involve a phase or part that concerns the identification of relevant stakeholders, such as individuals or groups affected by those impacts. In principle, the framework for the impact assessment procedure can depend on the agent conducting the impact assessment procedure and the context or activity within which the AI system is or will be deployed, such as whether an AI system is used for decision-making in the public or private sector. As noted previously, many researchers and organizations have encouraged the adoption of mandatory impact assessment processes for AI systems used in the public sector, for example, and often also private sector agents are encouraged to implement at least voluntary impact assessment processes as a means of self-regulation. Practical implementation has also initiated already with countries such as Canada implementing standardized structures and frameworks for impact assessment in the context of automated decision-making systems.

From the point of view of lawfulness, accountability and ethics, assessing the impact of an AI system that is (to be) deployed is desirable for many reasons. First, it expresses appropriate respect to citizens' and the general public's right to know what kinds of decision-making systems and algorithmic policies they are subjected to and what kinds of decisions impact their live [14]. However, this naturally necessitates that the findings of the impact assessment process are disclosed to affected individuals with due regard for both public transparency, on the one hand, and safety, on the other, disclosing only information that does not compromise confidentiality agreements or critical operations. Second, a standardized and well-defined impact assessment process can increase organizations' internal expertise and their capabilities with regard to evaluating AI systems or automated decision-making they develop or procure from third-parties. In other words, implementing an impact assessment process can help an organization predict and quickly identify ethical and legal issues and concerns which might affect the safety of the AI system or result in legal violations (e.g., discrimination). Third, impact assessment processes can in part ensure that the general public has a meaningful opportunity to evaluate, respond to and, whenever necessary, contest the use of an AI system in case the use of the system is not considered appropriate or socially responsible by affected communities.

Impact assessment processes nonetheless also involve risks. For example, a genuine concern is that, unless impact assessment processes are subjected to appropriate oversight, and unless their findings actually come to affect key decisions concerning AI system development and use, they can become ineffective and performative, thereby only legitimizing organizations' unethical conduct. For example, as there can be variance between different agents and communities in what exactly is considered to constitute a relevant impact or harm, the concepts and metrics that organizations might use for estimating any resulting harms can become "inappropriately distant from the harms experienced by people", thereby having the effect of rendering impact assessment processes actually into obstacles to "building the relationships required for effective accountability" [78]. In this sense, it is crucial to an effective process of impact assessment that it actually addresses the legitimate concerns affected communities might have.

Auditing algorithms

Algorithm auditing refers to a process where an AI system or a similar decision-making system is reviewed from th perspective of compliance, performance or fairness, for example. Algorithm audits are primarily technical in the sens that they involve reviewing and testing an AI system or inspecting technical documentation concerning the system . This does not mean that the review focuses on mere technical properties of the system (such as the performance o r speed of an algorithms). Rather, it means that any broader and possibly legally or morally relevant aspects or effects of the system, for example, are excluded from the assessment's scope. In comparison to impact assessment algorith m audits are thereby more restricted in their scope and purpose.

While algorithm auditing builds on many well-established fields and practices of auditing (a good example being safety audits in fields such as aviation technology), it is a is rather novel field. This novelty means that there are also various different frameworks available for algorithm auditing which differ in certain key dimensions. From a methodological perspective, for example, one can distinguish various approaches [8]. For example, in what is called a "code audit", the auditor obtains, reviews and analyzes the auditee's code. In the "direct scraping" methodology, the auditor collects data concerning the auditee's AI system directly via an application program interface (API) or other systematic query. "Sock puppet audits" involve an auditor collecting data about the auditee's system by creating computer programs or fake profiles impersonating users and using them to test the algorithm or the AI system. A "carrier puppet audit" is similar to the sock puppet audit method, but the impersonated users can have an effect on the algorithm or the AI system—and, by extension, on the end-users of the system. "Crowdsourcing" refers to a practice where the auditor collects data concerning the auditee's AI system by hiring end-users to test the system. Beyond these methodological differences, there are also other dimensions in which frameworks for algorithm audits differ. For example, algorithm audits can be either mandated or voluntary. The auditor can be either an internal or external agent from the auditee's organization, or the auditor might be appointed by a system developer, operator or a regulator, for instance. Algorithm audits can be conducted independently of the auditee's help, but there are also "collaborative audits" where auditors work together with product developers to identify problems with a software product [120].

Consensus on the exact methodology and framework for algorithm audits is yet to develop. Open questions and challenges relate to how algorithm audits ought to be conducted and by whom, what standards should guide them, and how they can be made effective in terms of leading to concrete actions. These issues are briefly discussed on the following page, and they should also be recognized and addressed at a regulatory level as informed by high-quality research. In any case, we suggest that algorithm audits constitute a central practice that should be implemented to ensure responsible use of AI systems. Not only do they provide a way to "peek inside the black box" but, as we will suggest, they can also complement impact assessment processes by identifying and mitigating risks and negative effects that are traceable to the AI system specifically. We will next briefly consider some open questions that concern the specific structure and effectiveness of algorithm audits.

A recent report by the Digital Regulation Cooperation Forum describes the various challenges looming within the current landscape of algorithm auditing [29]. To start, algorithm audits lack rules and standards in most sectors, especially those that are less regulated. Auditors can also enter the algorithm auditing market without assurance of the quality of their audits. Generally, there is also inconsistency and a lack of clarity pertaining to what algorithm audits should focus on. In practice, the effectiveness of algorithm audits is also undermined by various factors, such as prominent auditors (including researchers and non-profit organizations) lacking access to AI systems. Developers or controllers of AI systems can be reluctant to co-operate with auditors or they might provide the auditors inconsistent documentation, for instance. Auditors might also face risks of being legally sanctioned for collecting information about AI systems via "scraping" or probing them. Finally, a very concrete and pressing issue is that algorithm audits tend to lead to insufficient action, "particularly because the people affected have no means of receiving redress" [29]. Another reason, however, is that algorithm auditing can be "hijacked" by the audited party. For example, as sociologist Mona Sloane has argued, there is genuine risk that auditors' recommendations for improving audited algorithms are simply not taken into consideration and that the audit itself as a performative process is simply used to legitimize harmful technologies [100].

While the previously mentioned problems must be recognized and addressed, our general recommendation is that algorithm audits should nonetheless be made mandatory for all AI systems that are used—or are expected to be used—in high-stakes decision-making contexts, such as education or criminal justice contexts. Any mandated audits should be conducted by authorized external parties following best practices and applicable standards, including sector – or industry-specific standards when applicable. However, we note that algorithm audits should ideally be conducted also for the purpose of self-regulation in other contexts, for example, as a part of pro-ethical and responsible AI practices. Voluntary audits can and should also harness and leverage the capabilities of suitable external or third parties. For example, the research community, social activists, NGO's and the public could be incentivized to participate in so-called "bug bounties" or other activities that are organized to allow external parties to audit a given AI system [61].

Recommendations on impact assessment and auditing

We recommend that a requirement to assess the impact resulting from (or traceable to) the development and deployment of an AI system be included into four types of impact assessment procedures, and that they include audits of applied algorithms (or AI systems more broadly). This combined requirement would constitute an applicable, feasible and desirable mechanism for promoting algorithmic accountability. A well-structured impact assessment promotes compliance and safeguards individuals' rights and freedoms without imposing undue burdens on the assessor or party under assessment (such as developers and operators). Including an algorithm audit to impact assessment procedures is important to identify computing-related risks and effects. However, algorithm audits are also highly desirable for independent reasons, and they should be implemented to complement pre-existing processes of quality and risk management, for example, and to monitor system use for potential issues that arise in changing use-environments.

Impact assessment should be conducted prior to system deployment and regularly during use-phases (such as on annual basis). As there are different frameworks for impact assessment currently—both AI-related and others—we emphasize that any mandated process of impact assessment should leverage established and well-proven frameworks. In particular, for AI systems that are used or are envisioned to be used in high-stakes decision-making contexts, we consider frameworks for human rights impact assessment (HRIA) and fundamental rights impact assessment (FRAIA) to provide a well-established structure for assessing the impact of those systems. Both civil society organizations and researchers focused on AI and impact assessment have already proposed HRIA as an applicable framework in this respect, with researchers also having proposed a concrete structure for the HRIA in the context of AI system development and use [74, 75, 76]. We encourage responsible agents to utilize existing resources on HRIA for the purposes of impact assessment in the presently discussed context. However, we note that applicable alternatives for impact assessment frameworks might also include existing frameworks for algorithmic impact assessment (AIA) provided that they are customized to explicitly include an assessment of AI systems' impact on human rights and fundamental freedoms. Importantly, it should also be noted that responsible agents may be subject to independent requirements to identify and evaluate the impact of their policies, products or services. For example, they may be required to assess environmental impact, equality impact, and data protection impact of their business projects, policies, or products. Recognizing this possibility, we maintain that an assessment of the impact of developing and/or using AI systems should be included into any applicable and independently required assessments of impact. This includes cases where an AI system does not strictly speaking generate or result in some relevant impact but, rather, mediates that relevant impact—prominent examples including cases where an automated system is used to support human decision-makers executing a policy or decision-making process that is assessed, or when a product has an AI-based component, for example. The following page describes our proposal for a general framework for impact assessment in the context of AI systems. Below we also provide a description of the generic structure that impact assessments (but also ethical reviews) traditionally follow. In what follows, we will also consider three types of impact assessment in a bit more depth: data protection impact assessment, equality impact assessment, and environmental impact assessment.

AI systems should undergo impact assessment prior to deployment and regularly during use

Policymakers should require that agents developing and/or using an AI system in a high-stakes decision-making context should conduct an assessment of the system's impact on human and/or fundamental rights and freedoms—for example, a human rights impact assessment (HRIA) or a fundamental rights impact assessment (FRAIA). If such an impact assessment procedure is not required by law, responsible agents should conduct one in case they develop or use an AI system in a high-stakes decision-making context.

- Human Rights Impact Assessment (HRIA). An HRIA is conducted to systematically identify, analyze and respond to effects that a project, plan or policy may have on individuals' human rights and their effective access to those rights. The use of AI systems in increasingly various areas of life—including high-stakes decision-making contexts such as education, housing and criminal justice—involves risks for violating natural persons' human rights or preventing their access to those rights.

- Fundamental rights impact assessment (FRAIA). An FRAIA is conducted to systematically identify, analyze and respond to effects that a project, plan or policy may have on individuals' fundamental rights and freedoms, including their effective access to those rights and freedoms. Fundamental rights as a framework for impact assessment bears in many ways a close relation to human rights impact assessment frameworks, but is more geared towards national legal frameworks (than the international human rights framework).

Importantly, responsible agents can be (for independent reasons) required by law to conduct other types of impact assessment (see below). We recommend that those assessments include an assessment of the impact resulting from the development or use of an AI system specifically in the context of the assessed project, plan or policy.

- Data Protection Impact Assessment (DPIA). A DPIA is conducted to systematically analyze, identify and minimize risks that a project or plan involves in relation to data protection. Data mining, machine learning and use of AI systems can involve significant risks from the perspective of data subjects' anonymity and privacy, including risks that relate to model security and their vulnerability against adversarial attacks used to infer sensitive information.

- Equality Impact Assessment (EQIA). An EQIA is conducted to systematically analyze, identify and minimize risks that a project, plan or policy involves from the perspective of equality and affected individuals' rights to non-discrimination. Data mining, machine learning and use of AI systems can involve significant risks for violating decision subjects' rights to non-discrimination in case model training leads to the inheritance or amplification of bias that can lead to systematic differences system performance or expected outcomes across different protected groups (such as gender or ethnic groups).

- Environmental Impact Assessment (EIA). An EIA is conducted to systematically analyze, identify and minimize risks that a project, plan or policy involves from the perspective of environmental impact and sustainability. AI systems can have negative effects on the environment, especially when systems employ large models with extensive model training processes and massive datasets, or when they are run frequently to generate outputs. Software products also require hardware and infrastructure that consume energy and natural resources, and which result in CO_2 emissions.

PRO-ETHICAL AI PRACTICES:
Questions and prompts to guide impact assessment

STEP 1. Who or what is impacted by the AI system?

Who or what is directly affected?
E.g., natural persons, legal persons, groups or communities, areas and environments, physical objects or infrastructure.

Who or what is indirectly affected?
E.g., individuals' families, recreational opportunities, local market, local wildlife, relations between communities.

STEP 2. What is the type of impact or risk in question?

What fundamental value, legal norm, regulation or context-specific value is affected?
E.g., well-being, happiness, autonomy, freedom, dignity, lawfulness, non-discrimination, transparency.

What individual rights or freedoms are affected?
E.g., right to non-discrimination, right to education, right to housing, freedom of expression, freedom of movement.

STEP 3. What is the scope and severity of the impact?

What is the scale of the identified impacts or risks?
E.g., the problem affects only a few persons, the problem affects entire communities, the impact concerns everyone.

What is the degree of the identified impacts or risks?
E.g., the issue has only minor effects, the impact is moderate, there is a significant risk for some affected individuals.

What is the probability of each identified impact or risk?
E.g., the effects are highly probable, the risk is unlikely to realize, there is a medium-probability that the outcome occurs.

STEP 4. What is the overall risk or impact?

The evaluator should produce an overall assessment which orders identified risks and impacts in terms of three factors:

1) The scale or scope of the impact or risk;
2) The degree or severity of the impact or risk;
3) The probability or likelihood of the impact or risk.

The evaluator can use categories such as "low", "medium", "high" and "very high".

STEP 5. How should responsible agents and/or duty-bearers respond to identified impacts and risks?

Who is responsible for addressing identified impacts or risks?
E.g., who are the duty-bearers? Who are best equipped to address adverse effects? Who should be consulted?

What is the urgency of addressing identified impacts or risks?
E.g., an identified risk is low-priority, an identified adverse impact should be addressed immediately.

Is the identified issue or problem manageable?
E.g., can the problem be broken down to smaller pieces? Should the response focus on prevention or compensation?

How should the identified impact or risk be addressed?
E.g., are there robust and effective ways of responding? Is there a safe and lawful way of mitigating negative impact?

☑ Data Protection Impact Assessment (DPIA)

Data Protection Impact Assessment (DPIA) is required in many cases and contexts, including most contexts that we have called "high-stakes" decision-making contexts. DPIA requires an assessment of the envisaged information processing activities when dealing with personal data, including also an assessment of the necessity and proportionality of relevant operations in relation to the explicated purposes of processing and any expected or foreseeable risks resulting from those operations. A DPIA also covers the measures envisaged to address any identified risks, such as risks related to data subjects' privacy and the fairness of data processing. The aim of a DPIA is to establish whether any remaining or residual risk resulting from information processing is justified and acceptable given the circumstances and use-context in question. Overall, the DPIA is meant to help data controllers ensure, document, and demonstrate that they are in compliance with data protection regulations. We recommend that responsible agents utilize existing research and available resources on data protection in these contexts, such as the UK Information Commissioner's guidance on AI and data protection and their AI and Data Protection Risk Toolkit when and where applicable [56].

In what follows, we will focus on some relevant risks that arise—or are more likely to arise—in the context of machine learning applications and projects involving "data mining" specifically. Roughly speaking, machine learning can be noted to exacerbate privacy concerns due to at least three reasons: First, machine learning commonly relies on extensive amounts of data to work well, and hence agents may be at risk of violating individuals' privacy when collecting, curating, or transferring data. Particularly salient risks here concern the unlawful or morally unacceptable collection, distribution, or disclosure of personal data, where personal data is understood as data that concerns a natural person, or which can be used to identify a natural person. Second, analysis of massive amounts of data can lead to violations of privacy because the agent may, by using machine learning, inadvertently infer sensitive information about individuals or groups. Third, machine learning models are susceptible to attacks by adversaries and malicious agents who might wish to (re-)identify specific individuals, or to gain valuable information about model internals. We will discuss some general ways to manage and respond to these risks. Regardless of the chosen approach, however, we emphasize that risks related to personal data are to be viewed as risks that concern individuals' fundamental rights. As the European Data Protection Board (EDPB) notes, the "GDPR […] treats personal data as a fundamental right inherent to a data subject and his/her dignity, and not as a commodity data subjects can trade away through a contract" adding also that "the fundamental rights of data subjects to privacy and the protection of their personal data override, as a rule, a controller's economic interests" [36]. Consequently, a DPIA should, for the relevant parts, understood also as an assessment of whether and how data collection, processing, and distribution affects data subjects' fundamental rights. We also note that DPIAs should take into account the privacy norms of specific use-contexts, viewing privacy through the lens of "contextual integrity" where privacy is understood as a context-sensitive value, the protection and realization of which requires considering what kinds of roles, actions, practices, and normative expectations characterize informational contexts and domains [86].

A PERSPECTIVE:
Adopting a broader lens to privacy

The privacy discourse around AI is often framed quite narrowly. Privacy is considered to primarily questions related to information confidentiality, to the collection and processing of personal and/or sensitive information, to the possibility of (re-)identifying individuals in data, and to issues related to controlling data. While all of these questions are important in their own right, they do not exhaust the scope of possible issues that may arise from the perspective of privacy. On the one hand, existing approaches to preserving individuals' privacy do not necessarily capture social meanings of privacy—namely, what it means to respect the privacy of a person. On the other hand, existing protections on privacy are continuously undermined by practices that are prevalently used in the industry and many business contexts to circumvent privacy regulations and legal norms. For example, so-called 'dark patterns' are widely used to sidestep requirements to gain meaningful consent for data collection from individuals, and to condition and manipulate users to provide access to their data [35]. Similarly, intrusive surveillance is becoming increasingly more common in publicly accessible spaces, even though it often stands in practical tension with the notion of "meaningful consent". These issues highlight that there is a need to broaden the lens of privacy in the context of AI.

AI systems should be built and used in a manner that respects what Helen Nissenbaum terms "privacy as contextual integrity" [86]. Contextual integrity implies that an understanding of both (1) "what information about persons is appropriate, or fitting, to reveal in a particular context" and (2) whether the "distribution, or flow, [of information] respects contextual norms of information flow" (e.g., confidentiality, free choice, discretion, entitlement, and obligation) is required to come to a judgment about whether distributing or disclosing some specific piece of information is morally permissible or impermissible [86]. In other words, contextual integrity means, for example, that a given piece of information may be appropriate to disclose or distribute in one context but impermissible to disclose or distribute in another. An individual might be comfortable disclosing their pregnancy to a friend or family-member, for instance, but not to a recruiter during a job application process. This might be because the employer is expected to refrain from relying on information about candidates' health conditions (e.g., due to the possibility of health-based discrimination) or because the individual does not simply consider their pregnancy relevant to the context of applying for a job. An important point here is that norms and normative expectations concerning privacy are context-dependent, and thereby the question of whether a given way of disclosing or distributing information is morally permissible or not cannot be determined without considering contextual factors. Contextual integrity as a first step towards a broader understanding of privacy urges responsible agents to consider individuals' and communities' social and normative expectations regarding the collection and distribution of information, and to acknowledge the contextuality of such expectations and other relevant factors. In addition to protecting "private" or "personal" data, however, responsible agents could actively develop ways to empower and enable data subjects to draw their own boundaries between what they consider acceptable and unacceptable purposes of data collection and information processing.

Once we recognize how norms and expectations concerning privacy are shaped by contextual factors, we can also acknowledge that the collection and use of information can also affect groups (and not just individuals). "Group privacy" captures the idea that collectives, such as families, demographic groups, vocational groups, even sports teams, might have collective interests (perhaps even rights) regarding privacy [71]. For example, identifying patterns of behavior within certain demographic groups or parts of a population may violate normative expectations of privacy, but the use of inferences drawn about group-level behavior can also expose those groups to new kinds of harms (e,g., when statistical patterns related to consumer behavior are used for targeted advertising).

Together, contextual integrity and group privacy provide two important conceptual lenses through which privacy can be evaluated in te context of AI system development and use. On the one hand, contextual integrity calls responsible agents to ask not only whether a given set of data can be collected and how, but whether it is appropriate given the normative expectations that people have regarding data collection and processing in the use-context of the system and the domain where it operates. The notion of group privacy, on the other hand, urges responsible agents to consider not only risks that arise from the perspective of individual data subjects but also whether the inferences drawn from that data could lead to disproportionate or unnecessary harm towards a given group, for instance.

Technical privacy and security risks related to the development and/or use of AI-based applications can be distinguished on four key levels, as is shown below. These risk do not exhaust the possible scope of privacy-related ethical concerns in the context of AI system development and use, however—responsible agents will also have to consider infrastructure– and user-related risks as well as broader lenses to privacy (see above).

TABLE. Four key areas of data protection and security risks in AI systems

1. Training and evaluation data. Sensitive information could be disclosed in, or inferred from, the applied training and evaluation datasets. For example, an adversary could try to infer whether specific individuals are included in the training data, or seek to reverse-engineer the training data.
2. Input data. Sensitive information could be disclosed in, or inferred from, the input data fed into the AI system. For example, an adversary might try to observe the input data fed into a machine learning algorithm by the system's operators or users of an online platform.
3. Model or algorithm. Sensitive information could be disclosed in, or inferred from, the predictive model. For example, some machine learning models include examples from the training data by default, but an adversary could also perform privacy attacks to infer sensitive information from a given model (see [110]).
4. Output data. Sensitive information could be disclosed in, or inferred from, the output data generated by the AI system. For example, an adversary could prompt the model in ways that lead it to disclose sensitive information.

Some relevant privacy risks are not particularly specific or unique to AI and machine learning. Most projects involving data collection, storage, transfer, or publishing, for example, involve risks for violating individuals' privacy and breaching information security. There are risks related to data collection ranging from non-consensual data extraction (e.g., when developers scrape images of copyrighted material or human faces from websites) to otherwise intrusive or unacceptable interference with individuals' private space (e.g., tracking people in public spaces without their knowledge). We discuss some of these risks later in this document in connection to privacy notifications and consent management. Risks related to data storage, publishing, and transfer include the many potential ways in which sensitive information, such as information that can be used to identify individuals, might be inadvertently leaked. Both types of risks should be actively identified and evaluated with different methods, including by using technical privacy metrics to quantify relevant privacy risks [116]. Ways to manage and mitigate any identified risks might include, for example, ensuring compliance with data protection principles and using privacy-enhancing technologies (PETs). The principle of data minimization, for example, requires that personal data is used only to the extent that is sufficient to properly fulfill the stated purpose of processing, in ways that exhibit a rational link to that purpose, and only to the extent that is necessary. In other cases, responsible agents could apply PETs (e.g., anonymization, pseudonymization, data perturbation, secure multi-party computation, and differential privacy technique) to protect sensitive data.

There are risks that are rather unique to AI-based applications, however, due to the nature of machine learning and the models those applications use: On the one hand, developers, operators, and procurers of machine learning models need to be aware of the possibility that the trained model itself may contain personal or otherwise sensitive data. This is because some machine learning models contain parts of the training data by design, as it were, prominent examples including k-nearest neighbors and support vector machines. On the other hand, so-called "model privacy attacks" can be used by adversaries or malicious agents to infer sensitive information even from other types of models. In particular, attackers may attempt to reverse-engineer the training data or parts of it by using the model's outputs, possibly gaining access to identifiers or data that can be otherwise used to identify an individual. There are different types of attacks, notably. In model inversion attacks, for example, attackers estimate the training data by using information about the model, whereas in membership inference attacks, they will try to estimate whether a specific individual is included in the training data (see [110]). Privacy attacks can be distinguished into "white box attacks," where the attacker has knowledge about the model (e.g., they have access to model parameters), and "black box attacks," where the attacker can only observe the output (e.g., by prompting the model). In the presently discussed context, model attacks constitute a specific area of risk that should be addressed in the context of the DPIA. Both of the previously mentioned types of risks should be mitigated with both technical methods—such as using privacy-preserving data mining methods [2]—and operational safeguards—such as establishing clear protocols regarding who is allowed to access personal data and when.

Responsible agents should also note that many privacy risks will have to be assessed also in connection to the use of explanation extraction methods, which will be discussed below. Generally speaking, explanation extraction tools—also called "explainable AI" methods—are tools that are designed to provide different kinds of information about machine learning models or specific outputs that they generate [81]. The risk here is that the use of explanation extraction can also lead to disclosure of sensitive information, or that an "explainable" model might be used by adversaries and attackers (or other malicious agents) to infer such information. Whether the use of explanation extraction tools involves these types of risks depends largely on the applied method and the explanations it is used to generate, however. The key point here is that, while explainability and transparency constitute highly desirable goals for AI system design, responsible agents should always weigh the benefits and risks of achieving transparency against other legal and ethical requirements that concern model security, confidentiality, and the protection of data subjects' privacy.

On the following page, we provide some general recommendations on how responsible agents should identify, manage and respond to the variety of risks related to privacy and security throughout the lifecycle of an AI system in question. While the list is not exhaustive, it provides some actionable guidance from both a technical and operational perspective. Whatever methods and practices responsible agents adopt to ensure that AI systems are designed to protect affected individuals' privacy, those methods and practices should have as their central aim the protection of data subjects' fundamental rights, and the promotion of affected stakeholders' privacy, including both individual natural persons and groups.

CONCEPT / DESIGN

DEVELOPMENT PHASE

USE-PHASES

Build (and encourage the development of) AI systems that protect and enhance individuals' privacy

- Build AI for the purpose of protecting individuals' digital rights and privacy
- Prioritize projects and applications that minimize the need for personal data

Follow "privacy-by-design" principles

- Follow data protection principles and conduct a careful DPIA when necessary
- Adopt a proactive and preventive (as opposed to a reactive) approach to privacy and data protection
- Establish a plan for end-to-end security throughout the envisaged system's lifecycle

Align AI systems and data governance with contextual norms of privacy specific to the use-context or domain

- Identify contextual norms and normative expectations concerning privacy in the envisaged use-context
- Identify and respect stakeholders' rights and interests regarding the management of their data

☑ Equality Impact Assessment (EQIA)

An abundant research literature has demonstrated that data mining, machine learning, and the use of AI systems involves significant risks for discrimination and reproduction—even exacerbation—of substantive inequalities due to AI systems' inheriting or amplifying different types of biases that are embedded into to the applied sets of training and validation data [77]. A central concern relates to the representativeness of training datasets, a major risk being that the data used to train AI systems are not representative of human diversity and different demographic groups. The distribution of "ground truth" labels, which serve as proxies for true outcomes in prediction tasks, can also reflect and track pre-existing societal inequalities or result from human errors in data labeling, for example. The table on the following page describes some of the key types and sources of bias. Regardless of their sources, however, biases that are inherited and reproduced through machine learning can lead to systematic differences in whether and how well the AI system works when it is presented with profiles of members of different demographic or phenotypic groups, for example, or in how the (expected) benefits and burdens are distributed between different legally protected groups as a result of processing. Differences in AI system performance and outcome distributions can have legally and morally significant outcomes in real-life contexts. In 2019, for example, a study demonstrated that an algorithm that was used by hospitals and insurance companies in the U.S. to identify patients that would benefit from access to specialized care contained significant racial bias, consequently resulting in racial disparities in access [90].

Our recommendation is that an assessment of potential bias against legally protected groups should be included into EQIAs, which decision-makers should be required to conduct in high-stakes decision-making contexts. An EQIA is conducted to systematically analyze, identify and minimize risks that a project, plan or policy involves from the perspective of equality and affected individuals' rights to non-discrimination. Given the risks AI systems pose in these respects, it is of fundamental importance that the relevant negative effects that AI systems have—or are expected to have—are also actively identified and assessed prior to deployment, as well as appropriately documented and mitigated, when necessary. A recent report provides a fruitful framework for EQIA for AI systems in the context of hiring and recruitment [55] which can nonetheless be applied beyond the specific context of hiring. First, responsible agents should determine whether there is a need to undertake an EQIA. If there is, a general description of the AI system, including the purpose of its use, should be provided. After this, the evaluator should conduct a risk assessment the purpose of which is to identify potential risks and effects regarding equality and non-discrimination that relate to the use of the AI system. During this stage, a technical audit of the AI system should be conducted to quantify the expected negative and positive effects of using the AI system. The technical audit should make use of existing auditing tools and applicable measures of model fairness, which will be discussed below [111]. The results of the technical audit are to be combined with the general assessment of the deployment-related risks and impacts concerning equality and non-discrimination, which are then documented with detailed descriptions and explanations of whether and how those risks and impacts are addressed by the responsible party. The process should generate an EQIA statement including, for example, a description of the technical changes made to the AI system to align it with requirements concerning compliance with equality and non-discrimination law.

TABLE. Sources and types of bias throughout the lifecycle of an AI system
See also [77].

	TYPE / SOURCE OF BIAS	ILLUSTRATIVE EXAMPLES
PROBLEM	**Problem formulation bias** The way the computational problem (or objective function) has been formulated disadvantages some individuals by default.	An AI system is designed to predict and prevent undue access to a public benefit or service to which individuals have a right (as opposed to promote individuals' access to crucial benefits and services they are entitled to).
DATA AND FEATURES	**Sample bias** A trait, behavior or group is over– or underrepresented in the data sample, or represented disproportionately in relation to an actual or ideal baseline.	A dataset containing human faces which is used to train a facial recognition algorithm includes less elderly people due to biased sampling or self-selection. A dataset containing excerpts of human voices lacks training examples from persons with speech impairments entirely.
	Label bias Values of the target attribute are over– or underrepresented in some subpopulation included in the data sample, or represented disproportionately in relation to an actual or ideal baseline.	The prevalence of a disease, the presence of which a decision-maker wants to predict based on patients' profiles, is different among subpopulations, such as gender groups. A dataset used to train an algorithm for determining access to a learning group in an educational context lacks "ground truth" labels for false negatives, i.e., previous false rejections.
	Feature bias An included data category or model feature systematically disadvantages members of some group (potentially due to sample or label bias).	An assessment algorithm used for recruitment purposes includes language proficiency as a model feature, thereby systematically disadvantaging individuals with worse language skills. Notably, language proficiency functions as a proxy for immigrant status in the model as well, placing immigrant applicants into a position of disadvantage.
MODEL / ALGORITHM	**Algorithmic bias** By optimizing the pre-established cost or reward function, a trained model or algorithm mitigates or amplifies existing bias, such as data bias or societal bias.	A predictive algorithm used to identify potential locations of crime is optimized for predictive accuracy in past arrest data. This consequently leads to increased police presence in alleged high-crime locations, leading to more arrests as a result.
USERS AND APPLICATION OF THE SYSTEM	**User bias (including automation bias)** A user or operator of the AI system interprets, accepts, or rejects system outputs with systematic prejudice or bias.	An operator of an AI system does not know exactly what the output represents or means, and thereby takes outputs inconsistently into account when making decisions. An operator of an AI system exhibits a tendency to reject outputs more easily when they assume the profile belongs to a decision-subjects belonging to a minority group.
	Application bias (or deployment bias) The AI system is applied or deployed with systematic prejudice or bias, targeting specific individuals or a specific group.	An AI system is used to scrutinize only members of a given ethnic group. An AI system can be accessed only by members of a given vocational group. Members of other vocational groups are denied access, or accessing the system is disproportionately burdensome or difficult.
	Compounding bias and feedback-loops Multiple rounds of using the AI system amplifies negative outcomes for some individuals or groups, possibly due to interconnectedness between different decision-making systems' outputs.	A decision with negative consequences that is generated by an AI system for an individual leads to further similar decisions down the line, because a "negative spiral" is created. (See also example from predictive policing above.)
	Emergent bias By learning "online" during use, the AI system learns new proxies for group-membership (e.g., gender).	A recommender system used on an online shopping platform discovers a proxy feature for gender by learning from users' consumption habits. A user group learns that a facial recognition algorithm does not work for masked faces. The algorithm comes to exhibit a lower rate of correct recognitions for people who wear masks, accordingly.

An EQIA should make use of existing methodologies—both quantitative and qualitative—designed for the detection of unlawful and morally unacceptable biases and for evaluating model fairness (or "algorithmic fairness"). We will discuss these methodologies in more depth below, noting that they can also be utilized as evaluation and risk management methods in other pro-ethical AI practices that are implemented to ensure compliance with principles of ethics throughout the AI system's lifecycle. The primary of aim of employing such methodologies in the context of the technical auditing part of the EQIA is to identify whether the assessed AI system exhibits—or will foreseeably exhibit—direct or indirect discrimination against individuals based on legally protected or otherwise sensitive characteristics, including ones defined intersectionally. Fairness metrics can be also used to discover (un)desirable properties and tendencies of machine learning models, for example, and to identify trade-offs between them. However, it is important to emphasize here that none of these methodologies or metrics can provide conclusive evidence of the presence or absence of discrimination. They can only provide preliminary evidence in this respect, and do so only under strict assumptions concerning, for example, contextual factors and the integrity of the data pipeline.

The auditing party should evaluate both applied datasets (including training and validation datasets) and model (including those pre-trained by third-parties and those trained in-house) for discriminatory bias and other "algorithmic harms" using statistical, similarity-based, causal intervention-based, and/or qualitative methods of analysis. For example, fairness metrics are designed to capture (expected) discrimination in the machine learning model and thus prove a valuable tool for model developers and auditors. There are tens of metrics, however, each of which operationalizes a specific conception of "bias". Most fairness metrics are designed for specific models, moreover, most being geared towards supervised learning contexts where one requires access to "ground truth" labels (such as individuals' actual outcomes in addition to their profiles). For identifying applicable methodologies and metrics, we recommend utilizing reviews and surveys of bias detection and mitigation methods [77, 111] as well as research literature on their application in compliance with legal and moral norms [114, 115]. As a general rule of thumb, one has to first determine which (types of) metrics are applicable given the applied data and models. Second, one has to identify what equality duties and non-discrimination norms apply to them in the use-context of the system (e.g., formal versus substantive equality). Lastly, one has determine which metrics are best aligned with those duties and norms. Fairness metrics that quantify disparities in individuals and groups' treatment—such as Individual Fairness [32] and Calibration Between Groups [22, 62]—are good choices for identifying salient risks for prima facie direct discrimination. For evaluating risks of indirect discrimination, it is recommendable to use metrics that are able to capture whether a given group is systematically disadvantaged in relation to another. Statistical metrics—including but not limited to Conditional Demographic Disparity [114] and Equalized Odds [48]—can be used to identify different types of biases in these respects. To assist in bias detection, we have also listed some prompts on the following page that can provide initial guidance on how to detect different types of biases ranging from disparities in groups' representation in datasets (e.g., underrepresentation of minority groups) to harmful or anxiety-inducing items (e.g., slurs in text data). To mitigate risks for direct and indirect discrimination, for example, legally protected attributes and proxy features which strongly correlate with such attributes should be removed from the data or the model, unless there is a valid legal and moral justification for including them.

PRO-ETHICAL AI PRACTICES:
Questions and prompts to guide bias detection

STEP 1. Evaluate the formulated problem and/or core objective for bias.

- Has the problem been formulated in a way that reflects affected stakeholders' interests equally and proportionally?
- Has the core task to be executed by the AI system chosen in a way that leads to the most equitable outcomes?
- Are applicable legal obligations (e.g., positive equality duties) reflected in the chosen approach to information processing?

If the answer to any of the above is "no", reformulating the approach to processing and/or automation should be considered.

STEP 2. Evaluate applied datasets and model features for bias.

- Does the applied dataset include data on protected classes/groups or otherwise sensitive attributes?
- Does the applied dataset include information that could indirectly encode information about such classes, traits or groups?
- Does the applied dataset include information that could be considered derogatory, demeaning or discriminatory?
- Does the applied dataset lack representative examples from certain classes or groups?

- Does the model include protected classes as model features?
- Does the model include (complex sets of) features that might function as proxies for protected classes?
- Does the model include features that might track structural patterns of inequality or effects of historical discrimination?
- Does the model include a target variable or class that might receive an offensive, anxiety-inducing or derogatory value?

If the answer to any of the above is "yes", an overriding justification should be provided for including the data or feature.

STEP 3. Prompt and evaluate the trained model for bias.

Is the model non-comparatively fair by recognizing what each individual is owed independently?

For example, does the model include all necessary predictors required to ensure the accuracy of individual predictions? Does it include all the factors and variables that warrant consideration for moral and legal reasons?

Does the model adhere to the principle of formal equality?

For example, are different items (such as individuals' data profiles) treated differently and similar ones similarly? Does the model generate identical outcomes to individuals or items that differ only in terms of their protected class? Do predictions take into account information on protected attributes? Does the distribution of predictions or outputs track the base-rates found in the training data, referring to the distribution of "ground truth" labels therein?

Does the model exhibit equality of accuracy between relevant comparison classes?

For example, does the model exhibit equal true positive (or true negative) rates across protected classes of items or individuals? Does the model generate equally relevant outputs (e.g., recommendations) regardless of protected status?

Does the model impose equal risk to members of different comparison classes?

For example, does the model exhibit equal false positive (or false negative) rates across protected classes of items or individuals? Is the ratio of false positives to false negatives equal between protected groups?

Does the model exhibit equality of outcomes?

For example, are items and/or individuals from different protected classes equally likely to enjoy a benefit or a burden? Are items or individuals from different protected classes represented in an output (class), such as a list of recommendations or AI-generated images of persons, with equal probability?

Does the model generate the greatest (expected) benefit to those items and/or individuals who are least advantaged?

For example, does the output distribution conform to the so-called maximin principle, which requires expected utility is maximized for those who are worst off? Does the model "boost" the scores of underrepresented items or individuals (where justifiable given legitimate aims to enact positive action)?

Mitigating "bias"

Fairness analyses may indicate the presence of bias and possible disparities in the data, model or output distribution. Identifying unacceptable types of bias in the data and the model is not enough. Bias also has to be eliminated or mitigated with due regard for the legal duties and obligations that responsible agents might have in the specific contexts regarding equality and non-discrimination. The question responsible agents should ask upon detecting bias is, in other words, whether the identified bias or disparity in question can be justified in light of legal and moral norms, taking into account the sources of those disparities and the mechanisms through which they arise. In case disparities or other forms of pejorative bias are found and deemed unacceptable, the agent should devise a plan for intervention. However, determining an appropriate approach to bias mitigation is not easy since preventing unacceptable discrimination or inequitable outcomes can require both technical and non-technical interventions, and those interventions can result in trade-offs, including a decrease in overall predictive accuracy, for instance. Alas, bias mitigation can also have undesirable short– and long-term effects if done haphazardly without considering how the AI system will interact with the affected population and context [98]. General approaches to bias mitigation are depicted on the following page.

On the one hand, the most accessible point of intervention can be the AI system itself, and in some cases the data pipeline. There are many technical approaches to bias mitigation available that can be applied to adjust the training data, to constrain the algorithm and the learning process, or to balance the resulting output distribution [77]. Machine learning practitioners and other individuals involved in the technical development of AI systems can benefit from existing software toolkits that package these methods, and which have been developed explicitly for the purpose of testing software for discrimination. Examples of popular toolkits include, for instance, IBM's AI Fairness 360, Aequitas, Pymetrics Audit AI, Google's What-If Tool, Fair Classification, FairLearn, Weights & Biases, and the more design-oriented Algorithmic Equity Toolkit. Crucially, however, the toolkits that are available for fairness testing can differ in the methodological assumptions they incorporate and the data that is required to use them (e.g., "ground truth" labels or data on protected attributes). Responsible agents should thereby familiarize themselves with different toolkits before choosing which one(s) to apply, confirm that they have access to the necessary data, and consult domain-experts to ensure that bias mitigation tools can be applied with care and due regard for safety.

On the other hand, technical interventions can be widely insufficient to address and mitigate complex and deep-rooted societal inequalities, which are also often intersectionally stratified. In the worst cases, so-called fairness-enhancing interventions on models can be "ineffective, inaccurate, and sometimes dangerously misguided" [98] precisely due to the inherent complexity and dynamic nature of socio-economic disparities which simple statistical metrics often cannot capture [37]. Non-technical interventions—such as political or social interventions, or choices to refrain from algorithmic processing and opt for more flexible, human-executed decision-making processes—can be more appropriate in such cases. In any case, intervention design for bias mitigation should be done with care and due regard for contextual factors to ensure bias is mitigated effectively and safely.

A PERSPECTIVE:
Bias identification and mitigation: The case of disability bias

Effective approaches to both identifying and mitigating bias in AI systems requires considerable care and domain-expertise. This means that the use of statistical metrics and bias mitigation techniques can be insufficient or ineffective as a means of preventing harmful or discriminatory outcomes unless done with care and informed by social sciences perspectives, for example.

A prominent example of the complexity of bias identification and mitigation can be found in the context of disability bias—or "ableist bias". The term disability bias is used to refer to systematic prejudice against persons with disabilities, differential treatment of individuals based on (perceived) disability, and treatment that systematically and disproportionately disadvantages persons with disabilities. Disability bias has been identified in many AI-based applications. A study on natural language processing systems found that passages of texts mentioning 'disability' were likely to be classified as toxic and rated negative in terms of sentiment [52]. Further studies on other applications areas—including biometric identification, gaze tracking, and emotion recognition in AI-supported hiring software—have shown that individuals with physical impairments can be at higher risk of receiving misclassifications [11, 118].

Identifying disability bias in datasets and machine learning models is a highly challenging and sensitive task, however, since norms related to (dis)ability are not static and "given"—rather, they are continuously socially constructed in the interaction between humans and their physical, social, and digital environments. As researchers of the AI Now Institute note, AI systems also play a crucial role here since they inherit, construct, and enforce standards of normalcy with the consequence that certain individuals are effectively rendered "outliers" located beyond the normal distribution of attributes that statistical models often assume [118]. They further note that, whereas disability bias should be actively mitigated to avoid discrimination, there are various other ethical and legal concerns at play which "affect disabled people in particular and acute ways, including issues of privacy, consent, and the high stakes of (mis)classification in the context of asymmetric power relationships (such as, for instance, between patient and doctor)" [118]. A salient risk is that "de-biasing" machine learning models might end up only legitimizing medical discourses around (dis)ability, for instance. An example mentioned in existing research concerns AI-based tools that provide assistance in diagnosing autism but meanwhile legitimize "the notion that the formal diagnostic route is the only legitimate one for autistic existences, in turn reinforcing the power that psychiatrists hold" [118]. Instead of "de-biasing" models which risk legitimizing and automating practices and conceptions that in fact require transformation, in other words, sometimes the best thing to do might be to rather improve, reform, or resist ableist and exclusionary practices (e.g., social and healthcare practices).

"De-biasing" a model can prove to be the right way to go, however, but this means that responsible agents will require data on persons with disabilities—perhaps even data on specific disabilities—to do so. However, persons with disabilities may have it in their interests not to disclose sensitive information about themselves—or may wish to do so strategically. Questions related to people's privacy and meaningful consent should not be forgotten, in other words, and the interest to mitigate model biases to ensure that they work for everyone needs to be weighed against affected individuals' legitimate interests for strategic disclosure and preservation of their privacy.

Given the contextual and socially constructed nature of (dis)ability, recognizing disability bias in AI systems will likely require experiential knowledge and insights that only persons with lived experience of disability can provide. This has a concrete and practical implication for designing and building accessible AI systems: design and development processes should involve consultation, collaboration, or co-design with persons with disabilities, caretakers and/or representatives from interest groups, for example. The slogan "design with, not for" captures that proactive actions of inclusion in these respects are necessary to take early on in the process of design and development: "It is no good waiting until the testing or evaluation stage to involve disabled people. This needs to happen as soon as design begins, preferably during the conceptualization of the product/service concept. What is needed is true co-design, where disabled people are part of the design team and the process of design. This should include a representative group of people with a diverse range of disabilities." [101]

☑ Environmental Impact Assessment (EIA)

As was discussed in relation to the principle of beneficence in section 2, AI systems can have a significant environmental impact which needs to be estimated and mitigated prior to deployment, when negative. To this end, we recommend that responsible agents include computing-related impacts to any Environmental Impact Assessment (EIA) independently required of them. Ideally, the environmental impact of AI systems—both traceable to their development and operational phases—would be assessed and mitigated regardless of context.

Generally speaking, an EIA consists in the identification and assessment of the environmental effects of a plan or project—including notably also its alternatives. Responsible agents will prepare an environmental report (or an EIA statement) and carry out consultations with relevant stakeholders. The findings of the assessment, including the results of consultations, are accounted for by the responsible agent during decision-making regarding the plan or project, and information regarding key decisions are disclosed to relevant parties (e.g., regulators). An EIA should in the present context also include the impact of developing or implementing an AI system, ideally covering the environmental impact of the entire AI system lifecycle. This has two key implications that should be explicated.

The first implication is that the EIA should take into account the impact of developing and/or manufacturing of the system, on the one hand, and the impact of running it, on the other [58]. For example, as the training phase of an AI system can result in large amounts of emissions, especially with larger models that have a massive number of parameters, an EIA focused merely on the so-called "operational emissions" of the system—namely, emissions resulting from running the system to generate outputs—will fail to generate a holistic picture of the system's overall environmental impact. There are also "embodied emissions", which result from the extraction of raw materials for building hardware, for instance, which will remain uncaptured by such an EIA. This is to say that responsible agents should estimate emissions and energy usage that can be traced to the utilized hardware and necessary infrastructure, and that they should monitor the system's energy usage and emissions also regularly throughout different phases of system use. The second key implication is that, to generate a holistic picture of the environmental impact of an AI system, the EIA process should cover emissions and other relevant effects resulting from what can be called "computing-related impacts", which cover operational and embodied emissions discussed above, on the one hand, and what can be called "application-related impacts", which refer the impacts resulting from the use of the system to optimize or execute existing processes and activities, on the other [58]. If one were to evaluate the environmental impact of an AI-based heating system for a building, for example, one would require information concerning both the energy consumption and emissions of the AI system itself, but also the expected effects of deploying the system (e.g., buildings' reduced energy consumption due to optimized heating). Of course, there can also be broader effects as well, which can be difficult to estimate due to their aggregated or complex nature, but which are nonetheless highly relevant from the perspective of assessing AI systems' sustainability [58].

The following page illustrates the variety of factors on which the environmental impact of AI systems depends.

PRO-ETHICAL AI PRACTICES:
Estimating the environmental impact of AI systems

IMPACT ON ENVIRONMENT DUE TO COMPUTING

Embodied impact
E.g., ore and plastic required to build hardware

Operational impact
E.g., energy consumption and emissions resulting from model training and model inference

The degree and scale of embodied and operational impacts depends on...

E.g., natural resources required for hardware, which results also in waste.

E.g., the amount and nature of collected and utilized data.

E.g., data centers and cloud-providers, which consume resources and energy.

E.g., the training process, model size and architecture, algorithm efficiency.

E.g., logistics and transportation requires material infrastructure and energy

E.g., component durability, maintenance, and the cooling capacity of hardware.

POSITIVE AND NEGATIVE IMPACT ON ENVIRONMENT DUE TO APPLYING THE AI SYSTEM

Optimizing an existing task or process
E.g., optimizing heating or lighting in buildings

\+ E.g., reduces energy consumption

− E.g., has a larger operational impact due to using a large machine learning model

Substituting an existing process or system
E.g., automating transportation

\+ E.g., more efficient driving routes in comparison to human drivers

− E.g., sustains or even increases reliance on private transportation as opposed to mass transit

Relevant impacts might range from micro-level effects (such as effects on consumer behavior) to broader effects (such as effects on how entire industries function at a systemic level).

An EIA should follow standardized methods and best practices for evaluating environmental impact. However, since estimating the environmental impact of AI systems—namely, specific types of software products and services—can be immensely difficult, we will here provide some theoretical resources and observations to assist in such efforts.

Recall the previously introduced distinction between "embodied" and "operational" emissions. We might note that estimating the former requires that the amount of natural resources and energy required to build and/or use the applied hardware and the adjacent digital and material infrastructure is estimated as accurately as possible. This serves to ensure that the resulting emissions can be quantified and mitigated—or, when necessary, compensated for—properly. On the one hand, the overall carbon footprint of an AI system is dependent on the hardware and components that comprise its material shell. Responsible agents should, notably, opt for efficient hardware that has good cooling capacities, for example, and prioritize renewable materials when possible, since the hardware used for computer systems will result eventually in waste (e.g., plastic). On the other hand, there are environmental costs which are dependent on factors external to the AI system itself, such as logistics and other infrastructure, the emissions of which are "embodied" by the hardware. This is why they also should be figured into the estimate of the environmental impact of the system.

An assessment of operational environmental impact—namely, the carbon footprint resulting from running the AI system—also depends on the previously mentioned factors. For example, the utilized infrastructure—ranging from cloud providers and data centers to electricity grids—can affect the overall impact of an AI system significantly. Data centers and cloud providers can differ in the extent to which they rely on renewable energy, for example, and in terms of whether they compensate for their carbon emissions. Responsible agents should prioritize infrastructure that runs exclusively or primarily on renewable energy. However, operational emissions are also dependent on technical factors, such as the chosen approach to model training (i.e., building the model) and model inference (i.e., generating outputs with the model). All computation is thereby not created equal in terms of its sustainability, in other words, which suggests that evaluating and comparing the energy-usage and carbon emissions traceable to not only data collection but also to model training and inference is highly important. For example, regarding hyperparameter search and tuning, it has been shown that grid search can be less efficient than alternatives, such as random search, but also costly in terms of carbon emissions [12, 65]. Quite intuitively, factors ranging from the size of the model as measured in the number of parameters in the model to the time it takes to train the model all make a difference since bigger and more powerful models require more resources—and, by extension, more energy—to train. Responsible agents should design their approach to be as energy-efficient and sustainable as possible, respectively, prioritizing sustainability-related in decisions regarding choices of hardware and modelling approaches.

On the following page, we have summarized our recommendations and some best practices based on existing research on sustainability in AI [58, 65, 109]. The recommendations cover a variety of actions ranging from measuring and disclosing AI systems' embodied and operational emissions to adjusting and designing supply chains to reduce emissions and waste. Hopefully, these recommendations can pinpoint some initial steps towards greener AI lifecycles.

CONCEPT / DESIGN

DEVELOPMENT PHASE

USE-PHASES

Build (and encourage the development of) AI systems that explicitly pursue valuable environmental goals

- Use AI to achieve SDGs related to climate action and environmental protection [95, 112]
- Prioritize data science projects with positive environmental impact

Adjust business model or organizational processes to avoid ML applications that contribute to climate change

- Do not distribute or procure software products that, for example, sustain or increase dependency in fossil fuels, consume extensive amounts of non-renewable energy, or which result in significant amounts of waste

Prioritize green supply chains, partners, providers and infrastructures

- Figure sustainability-related factors into decisions regarding affiliations, business partners, and infrastructures (e.g., using only energy that comes from low-carbon regions)
- Review cloud-providers and data centers' sustainability commitments, checking whether they have a Renewable Energy Certificate (REC), and opt for ones that use exclusively renewable energy

Include sustainability targets and thresholds into the AI ethics framework

- Establish a hard zero-emissions constraint on the operation of AI products and services—or, ideally, on organizational processes overall (e.g., 24/7 Carbon Free Energy)

PUBLIC TRANSPARENCY: EXAMPLES OF BEST PRACTICES

Transparency has many meanings and, in an instrumental sense, it can serve various functions throughout the lifecycle of an AI system [4]. On the one hand, machine learning systems can be "black boxes" in the sense that human users or decision-subjects cannot know or understand how the systems process information and how specific outputs are generated. Understood as a property of an AI system, "transparency"—or explainability, intelligibility or understandability—is thereby essential for safety in system use and for ensuring decision subjects' access to their rights. On the other hand, there can also be a lack of transparency in a broader sense—namely, concerning the purposes AI systems are used for, where and when, and by whom. Transparency and openness regarding these broader questions is central from the perspective of ensuring that "smart" software applications are used with due regard for public accountability and democratic legitimacy.

Transparency should be ensured in both senses of the word that were discussed above—both as an instrumentally valuable property of a technological artefact and as an instrumentally valuable property of the relation between different stakeholder groups. To this end, we propose transparency and documentation requirements that concern (1) AI systems, their components and functionalities at a technical level and (2) organizations developing or using them at a broader level of operations. Our recommendations are summarized in the table on the following page, and we will discuss them in more detail below. As a disclaimer, we might point out that the list of recommendations involves recommendations that have been proposed also in the European Union's draft AI act. We also note that the recommendations include recommendations regarding both desirable system properties and capabilities as well as proposed operations and pro-ethical AI practices.

The various types of requirements proposed on the following page are motivated by distinct sets of ethical and legal concerns. On the one hand, many of the recommended requirements and transparency practices are meant to ensure that responsible agents establish and evince public accountability by creating and maintaining clear audit trails. For example, agents should implement practices for documenting applied datasets and models as well as adopt practices of disclosure which seek to ensure appropriate public transparency throughout the lifecycle of their AI systems. On the other hand, some of the recommendations are motivated by specific issues relating to the complexity of AI systems, or specific risks related to their use. For example, given that AI systems possess increasing capabilities to fool, mislead or even deceive end-users, we consider it necessary to implement relevant safeguards such as interaction notifications and "transparency labels" that mitigate these risks. Other risks, which relate to the complexity and opacity of AI systems, should be mitigated by implementing appropriate methods of explanation extraction. However, the choice of an appropriate method for explaining AI systems' behavior or outputs is a delicate matter, as we will see.

RECOMMENDATIONS:
Transparency and documentation requirements

REGISTRATION REQUIREMENT FOR AI SYSTEMS USED IN HIGH-STAKES CONTEXTS

Any AI system deployed in a high-stakes decision-making context should be registered in an AI register accompanied by information concerning the party operating the system, the purpose of the system's use and any other contextually relevant information. The AI register should be accessible to the general public

PRIVACY NOTIFICATIONS AND TRANSPARENT CONSENT MANAGEMENT

Any AI system deployed in any context should adhere to applicable laws and regulations concerning data protection and privacy, including requirements concerning privacy notifications and consent management. Consent management platforms (CMPs) should adhere to standards of meaningful transparency and refrain from employing deceptive designs.

INTERACTION NOTIFICATIONS AND TRANSPARENCY LABELS

People should be notified when they are an interacting with an AI system. Depending on the context or circumstances of use, this notification should be directly provided for the person interacting with the system or it should be indirectly provided by making the information clearly and easily accessible. In addition, any AI system generating content such as synthetic images, text, audio or video should include a context-appropriate transparency label (such as a watermark, textual notification or icon) indicating the user that the content has been generated by an AI system.

ACTIVITY LOGGING CAPABILITIES

To create traceable audit trails, AI systems should be equipped with activity loggers (such as event data recorders) that maintain a log of who has accessed, maintained, operated, or otherwise interacted with the system and when.

EXPLANATION EXTRACTION METHODS

Any AI system used in high-stakes decision-making contexts should be equipped with a context-appropriate, reliable and effective explanation method. The method should be applied to make the behavior and/or output intelligible to relevant parties to ensure that affected individuals can access their rights and contest decisions, and to promote safety in system use.

DOCUMENTATION OF DATASETS AND MODELS

Datasets applied for purposes of training, evaluating or validating machine learning models should be documented. Documentation should include information about the dataset's origin or source, type and size, purpose of use, reliability and accuracy (including sources of possible errors and biases), possible legal or ethical risks related to its use, and other contextually relevant information.

Machine learning models and algorithms applied in an AI system should also be documented. Documentation should include information about the model's type and size, purpose of use, performance (including things such as accuracy, error rates and bias), and possible legal or ethical risks related to its use. Model documentation should also include information about methods used to improve the model's performance, transparency, security, fairness, and other contextually relevant information.

RESPONSIBLE COMMUNICATION

Organizations should commit to truthful, inclusive and responsible approaches to communicating about AI systems. This implies disclosing any benefits and limitations of developed or used systems truthfully with integrity and due regard for "downstream use" of the system. Communications should also use language that does not create or contribute to excessive "AI hype" or invoke harmful gendered, racialized or ableist stereotypes about the relation between "intelligence" and human attributes.

☑ Registering AI systems

We have noted previously that individuals and the general public deserve to know when and how they are affected by the use of AI systems, who is using those systems and for what purposes. Our recommendation is thereby the following: when used in public, semi-public or high-stakes decision-making contexts, AI systems be registered to a public database to inform the public of their use and to enable public examination of social acceptability and contestation, when necessary. Prominent examples of existing initiatives and registers include the AI register implemented by the City of Helsinki and the algoritmeregister implemented by the City of Amsterdam.

☑ Privacy notifications and transparent consent management

Especially when used in high-stakes decision-making contexts, AI systems tend to be data-intensive in that their development and use requires massive amounts of data. This introduces various risks from the perspective of individual and group privacy (such as loss of anonymity and harms that can follow from re-identification). However, beyond privacy risks related to datasets, salient problems and risks pertain also to the often ubiquitous and extensive data collection processes themselves. In data-intensive contexts, the perhaps most salient problem currently is that data collection practices tend to be exploitative and opaque. For example, in most cases, issues with regard to gaining individuals' meaningful consent for collecting data is not that organizations do not know that meaningful consent is required, or when and how it should be gained. The bigger problem is arguably the fact that organizations may either actively try to sidestep requirements related to consent or that they employ consent management platforms (CMPs) which are not designed to be accessible and intelligible. For example, research shows that a very significant number of CMPs implemented on websites do not meet the requirements stated in the GDPR [89]. The prevalence of so-called "dark patterns" on online platforms further suggests that business objectives and organizations' hunger for data tend to take priority over individuals' digital rights. Dark patterns are "interfaces and user experiences implemented on social media platforms that lead users into making unintended, unwilling and potentially harmful decisions regarding the processing of their personal data" and which "aim to influence users' behaviour and can hinder their ability to effectively protect their personal data and make conscious choices" [35]. For example, organizations may design CMPs or online platforms in ways that overload individuals' cognitive processing so that they get frustrated and quickly click on buttons that allow organizations' to access their data, or in ways that confuse users so that they do not what terms they have actually agreed to.

Transparent and meaningful consent management is necessary for protecting human autonomy, which is why emphasize that any AI system deployed in any context should adhere to applicable laws and regulations concerning data protection and privacy, including requirements concerning privacy notifications and consent management. CMPs should adhere to standards of meaningful transparency and refrain from employing deceptive designs which are arguably in direct tension with humans' cognitive autonomy and their right to informational self-determination [66].

☑ Interaction notifications and "transparency labels"

Humans should be able to know whether and when they are interacting with an AI system. This notion brings us to the following two recommendations. First, in line with the transparency obligations included in the European Union's draft AI regulation (Article 52), we recommend that humans should be notified when they are interacting with an AI system (such as a conversational agent or a chatbot). Depending on the context or circumstances of use, this notification should be directly provided to the person interacting with the system by showing a prompt or a message, for example, or indirectly provided to them by making the information clearly and easily accessible with reasonable effort. An exception to this requirement pertains to cases where the fact that the user or decision-subject is interacting with an AI system is entirely clear from circumstances or context. Second, we recommend that so-called "transparency labels" are applied to indicate when a given piece of content (such as a passage of text or an audio sample) is generated by an AI system. For example, any AI system generating content such as synthetic images, text, audio or video should include a context-appropriate transparency label which indicates to the user that the content in question has been generated by an AI system. An appropriate label might consists of a watermark, textual notification or icon, for example.

☑ Activity logging capabilities

Creating meaningful and transparent audit trails concerning the use and outcomes of AI system is a key part of algorithmic accountability. One should be able to determine who is responsible for a given decision or outcome, who is responsible for addressing or responding to a given problem that arises as a result, and who should be held to account in these respects. To promote these facets of responsibility and accountability in the use of AI systems, our recommendation is that AI systems should be equipped with activity loggers (such as event data recorders). Activity loggers should maintain a record of who has accessed, maintained, operated, or otherwise interacted with the AI system and when. Equipping AI systems with activity logging capabilities in especially important in high-stakes decision-making contexts due to concerns of liability, redress and remedy, for example, but implementing systems with such capabilities is also preferably in other contexts.

☑ Explanation extraction methods

Transparency and explainability have been at the center of discussions on AI ethics primarily due to the pressing problem of "black box" algorithms. Explainable AI methods [81] seek to address this issue by "opening the black box" or "peeking inside it". The following page provides a brief introduction to these topics, mentioning some relevant risks and approaches. Our general recommendation here is that responsible agents ought to select and implement explanation extraction methods with due regard for their truthfulness and accuracy, and stakeholders' informational interests and empirical facts about human psychology. In addition, moral and legal norms that regulate the use-context of the AI system should be taken into account together with concerns related to system safety, security and privacy.

A PERSPECTIVE:
Explanation extraction for explainable AI

Tr
interests.
algorithms —perhaps
even —understand their processing logic or the "reasoning" behind specific outputs. Complexity is introduced by the massive
amoun -dimensional
fe
an
constit
meaningful —after all, in a democratic society, justice
must n

In
of mac -called "explanation extraction" refers to the activity of explaining or generating information
either
latter "
an
al]. Some can be applied to all models, others can be used only with
specific
of
(RF

Importa
th
go
correlatio
Explanati
syste -of-distribution cases, for instance. Whether and what kinds of explanations one needs to extract from
th
explanati -makers using AI systems might require different information concerning model outputs than
de -experts may need
different
end-users will not find such "explanations" of an AI system's behavior satisfactory. Furthermore, explanations can be generated at different
times
v -case, one may have to implement one or more of these
ap -
context
explanati -based application, and what kinds of explanation extraction methods designers
an -context in which an AI system is applied can be governed
by -
critic -specific standards may partly regulate what information should be disclosed to users.

For these r

PRO-ETHICAL AI PRACTICES:
Questions and prompts to guide explanation extraction

STEP 1. Determine whether there is a need to extract explanations in the AI system's use-context.

- ☑ Is it necessary or desirable—for legal, ethical or pragmatic reasons—to extract explanations in the given use-context?
  E.g., are there legal norms or safety standards that require the operator to explain model behavior or outputs?
- ☑ What explanation extraction methods are available for the applied machine learning model or AI system?
  E.g., what methods are available and applicable for the applied model or inference technique?

STEP 2. Determine the ideal formal structure and semantic content of the explanation.

- ☑ What should be explained (i.e., the explanandum)?
  E.g., should the training data be explained? The model? A given output?
- ☑ What explains the explanandum (i.e., the explanans)?
  E.g., should the explanation provide a list of reasons, causes or correlates?
- ☑ What is the ideal format of the explanation shown to the system operator or decision-subject?
  E.g., should the explanation use sentences, graphs, timelines, or something else? How is the explanation represented?

STEP 3. Choose an appropriate explanation extraction method.

- ☑ Can the chosen method be applied to the model or system in question?
  E.g., does the method work for high-dimensional models? Can it be applied to the type of data the application uses?
- ☑ Can the chosen method be applied to achieve the central aim of explaining?
  E.g., does it help to justify or describe the output? Does it provide the operator more control over the system?
- ☑ Are the limitations and shortcomings of the chosen method known and accounted for?
  E.g., do system operators know when and how the explanations might go wrong or what they cannot disclose?

STEP 4. Evaluate a diverse sample of extracted explanations for mistakes and risks.

- ☑ Do the generated explanations satisfy general criteria for what make an explanation a good explanation?
  - Are the explanations coherent and internally consistent?
  - Are the explanations informative and sufficiently detailed?
  - Are the explanations consistent and non-contradictory?
  - Are the explanations practically useful and helpful given their intended purpose?
  - Are the explanations clear, intelligible, and understandable?
  - Are the explanations complete in that they lack informational gaps?
  - Are the explanations observable in that they truthfully describe what they are supposed to describe?
- ☑ Do the explanations disclose information they should not disclose?
  E.g., do explanations render the model vulnerable to privacy-attacks? Do they allow malicious users to "game the system"?
- ☑ Do the explanations actually satisfy stakeholders' informational interests adequately?
  E.g., have they been tested with actual users in real-life cases? Are the explanations intelligible to different users?
- ☑ Do the explanations actually satisfy stakeholders' informational interests adequately?
  E.g., have they been tested with actual users in real-life cases? Are the explanations intelligible to different users?
- ☑ Do the explanations handle uncertainty and out-of-distribution cases?
  E.g., does the quality or truthfulness of explanations change depending on the observed case or context?

☑ Documenting datasets and models

The practice of documenting important components and features of AI systems—including applied datasets [10, 42, 50] and models [6, 79], as well as approaches to explanation extraction [102], for example—is emerging as a central method of operationalizing values such as transparency and safety in the context of AI system design and development. Documentation is indeed desirable for many reasons. As a standardized and regular process, documentation can help technology developers identify technical, legal, and ethical issues pertaining to the AI system in question. As artefacts, documents concerning AI systems and their components also help archive information, and to distribute and disseminate it between different teams (e.g., development and compliance teams) and to different stakeholders (e.g., users and regulators). From the perspective of algorithmic accountability, documentation is also integral to the creation of audit trails which regulatory bodies or other authorized bodies can use to establish liability, for example, and which responsible developers themselves can disclose in order to increase transparency to the public and affected stakeholders.

Pro-ethical AI practices, documentation included, need to be standardized and explicit to ensure they are effective and successful. In this regard, documentation should be made a natural and mandatory part of the workflow within an organization (or a specific team) in the context of technology development and data science projects.

Two important questions pertain to documentation—namely, what needs to be documented and how? We will next consider these questions and provide some recommendations. We will first provide useful references to research where responsible agents can find templates and frameworks for documentation, after which we will examine different types of documentation and provide our own template for dataset and model documentation which has been adapted from existing templates. We have also previously emphasized that responsible agents should also document any impact assessment processes they have conducted, including their findings. In this regard, we note that the documentation practices discussed here can also be used to complement impact assessment processes and the technical audits that are conducted as a part of those processes.

What should be documented? There is no universal answer to this question. However, we might note that different use -contexts (e.g., sectors and industries) may have regulations or standardized practices which (partly) determine the content of relevant documents and deliverables. Furthermore, the content that should be documented depends on what the responsible agent is trying to achieve with documentation: On the one hand, if the purpose is to promote safe and informed "downstream" use of the AI system, it is likely that the downstream user should have access to documentation that they can use to promote appropriate and safe use. Relevant information might include, for example, the intended uses of the system (or model), the envisioned user-group, and any limitations pertaining to using the system (or model). In addition, the downstream user might benefit from model performance measures and technical details. On the other hand, if the purpose of documenting a model, for example, is to ensure that it can be evaluated for risks of discrimination by regulators, then the responsible agent should document everything that is technically, legally, and morally relevant to evaluating the system and to tracing the consequences of its use. Prominent examples might include risks and problems related to bias and performance. Documentation practices should, nonetheless, ensure that any confidential information or trade-secrets are not disclosed to the wrong people when access to documentation is provided to third-parties, for example, and that the disclosed information cannot be used to compromise safety-critical processes and system operations. This is especially important since in certain cases it might be appropriate to disclose model vulnerabilities to the user, whereas in other cases disclosing such vulnerabilities might be dangerous given the risk that malicious agents might use the information to attack the model.

The research literature on documentation in AI has notably focused mostly on documentation that concerns specific components or facets of AI systems—including applied datasets and models. The literature on dataset and model documentation can provide responsible agents useful and practical guidance regarding how to document key items in these respects, offering also applicable templates which responsible agents can either apply as such or customize to their own needs when necessary. For example, responsible agents might benefit from Datasheets for Datasets [42], Dataset Nutrition Labels [50], and Data statements [10] which have been created to structure the documentation of datasets, their motivations and intended uses, their composition and collection process, and any ethical concerns that relate to their content or use. For model documentation purposes, responsible agents might wish to use approaches such as Model Cards [79] or Factsheets for AI [6]. In addition, some documentation frameworks, which can be highly useful, are tailored towards specific legal and ethical dimensions of AI systems, an example being Explainability Factsheets used to assess and document approaches to explanation extraction [102]. Projects such as Partnership on AI's Annotation and Benchmarking on Understanding and Transparency of Machine Learning (ABOUT ML) are also actively producing resources to support meaningful documentation in AI.

The following pages contain our example templates which responsible agents can use for assistance in documenting datasets and models. The templates are largely adapted from the previously cited literature, although we have made adjustments to some parts of the templates. We stress that these templates are not presented as novel documentation frameworks but, rather, they draw on previously presented research to provide the reader illustrative examples of what responsible agents should ideally document throughout the lifecycle of their AI systems (and how).

PRO-ETHICAL AI PRACTICES:
Dataset documentation template with questions and prompts 1/2

[NAME OF ORGANIZATION]	[DATE OF DOCUMENT]
NAME OF PERSON OR TEAM RESPONSIBLE FOR DOCUMENTATION	[CONTACT INFORMATION]
[DATASET NAME AND VERSION]	[IDENTIFIER AND ACCESS POINT]

INTENDED USE OF THE DATASET

What is the primary use-case or intended application of the documented dataset?

PROPERTIES AND COMPOSITION OF THE DATASET

Data properties

- What kind of data does the dataset include? For example, does it include images, text, video or audio?
- Does the dataset include single-category data or multi-category data?

Dataset size

- What is the size of the dataset?
- Is the dataset a sample from a larger dataset?

Sampling

- How was the dataset sampled? Was it sampled randomly or in a targeted manner?
- Is the dataset representative of the larger reference population and how was this assessed?

Limitations and risks

- What are possible sources of errors, noise, redundancy, and bias in the dataset?
- What limitations does the dataset have in terms of representativeness and diversity?
- Does the dataset include potentially offensive or otherwise harmful items?
- Have items been removed or redacted from the dataset, or added into it? Why?
- Are there other concerns (e.g., legal, ethical, or safety-related) regarding the content of the dataset?

MAINTENANCE AND EXTERNAL SOURCES

Maintenance

- Is the dataset regularly updated? Who maintains and/or updates the dataset?
- Are older versions of the dataset maintained?
- Can users or other third parties contribute to the dataset? How?

External sources

- Does the dataset link to external sources?
- Are there official archival versions of the complete dataset?
- Are there any restrictions or fees that concern the transfer, storage or use the dataset?

DATA COLLECTION AND PREPARATION

Data collection

- How was the data collected and by whom?
- Does the dataset contain raw data, synthetically generated data and/or inferred data?
- Did collection involve human curation, self-reporting, or the use of hardware (sensors) or software (APIs), for example?

Pre-processing

- Did the data undergo any pre-processing procedures such as cleaning, labeling, discretization, bucketing or enrichment?
- Does the dataset contain targets or labels for data items?
- What are the targets or labels for, what information do they provide, and how were the data labeled?

Validation

- Has raw data been stored or saved in addition to the pre-processed data?
- How was the integrity, accuracy and validity of the data verified?
- How is the continuing representativeness, integrity, accuracy and validity of the data verified upon maintenance or updates?

DISTRIBUTION OF THE DATASET

- Is the dataset open access?
- Is the dataset distributed to third-party agents?
- Is the dataset distributed under a copyright or other intellectual property license?
- Do export controls or other specific regulatory restrictions apply to the dataset?
- Are there fees or other costs associated with distributing or accessing the dataset?
- Has the dataset been applied in existing systems or research?
- What kinds of tasks may the dataset be used in the future?
- What limitations or out-of-scope use-cases does the dataset have?

FURTHER NOTES

Are there any further remarks or notes on the dataset?

This dataset documentation template is adapted from existing research on dataset documentation with a few modifications [10, 42, 50].

PRO-ETHICAL AI PRACTICES:
Model documentation template with questions and prompts 1/2

NAME OF ORGANIZATION	DATE OF DOCUMENT
NAME OF PERSON OR TEAM RESPONSIBLE FOR DOCUMENTATION	CONTACT INFORMATION
MODEL NAME AND VERSION	IDENTIFIER AND ACCESS POINT

INTENDED USE OF THE MODEL

- What is the primary use-case or intended application of the model?
- What kinds of tasks is the model used for?
- Who are the envisioned primary users of the model?

DETAILS ABOUT THE MODEL

Model type and training

- What is the type of the model (e.g., classifier, neural network, convolutional neural network etc.)?
- What learning techniques and algorithms were used to train the model?
- What kind of an objective function has been built into the model?
- Was the learning process constrained somehow (e.g., by using fairness constraints)?
- Does the model learn "online" (i.e., continuously during use)?

Model inference

- What kind of input data does the model use?
- What kind of output does the model generate?
- How does the model generate the output (cf. model parameters)?
- What features does the model include?

EVALUATION AND METRICS

Performance

- What methods and metrics were applied to evaluate model performance?
- How does the model perform against applied metrics and benchmarks?
- What decision-thresholds were applied (e.g., threshold for placing an item into an output class)?
- What are sources of error in the model?
- What approaches were applied to address or reduce variance and/or uncertainty?

Ethical and safety evaluation

- What methods and metrics were applied to assess the model from ethical perspectives (e.g., privacy, fairness, diversity)?
- How does the model perform against the applied metrics and benchmarks?

FINDINGS OF QUANTITATIVE AND QUALITATIVE ANALYSES

Model performance analyses

E.g., illustrate model performance using a confusion matrix of model performance metrics.

Findings of fairness analyses

E.g., describe whether and how the predictive accuracy of the model varies across a set of unitary and intersectional comparison classes, such as classes of items specified by demographic or phenotypic attributes.

Findings of other ethical and safety analyses

E.g., describe the degree of representational diversity in a generative algorithm's output
E.g., describe how model explainability was assessed.

Describe also the limitations pertaining to evaluation methods or experimental setups.

DATASETS

- What datasets were used for model training?
- What datasets were used for model evaluation?
- What was the training/evaluation data split?
- Was the data pre-processed or supplemented (and if so, how)?

FURTHER REMARKS AND RECOMMENDATIONS

Limitations

- Are there noteworthy uses or applications for which the model cannot or should not be applied?
- Are there limitations that should be mentioned concerning the model, its performance, or its use?

Ethical considerations

- What ethical issues have been identified as pertaining to the model or its performance or behavior?
- Have the identified issues been addressed (and if so, how)?
- What recommendations can be given to model users to enable them to address the identified issues?

This model documentation template is adapted from existing research on model reporting with a few modifications [6, 79, 102].

☑ Responsible communication about AI

AI ethics research has identified many problems that pertain to the marketing of AI systems and the way in which the affordances and limitations are communicated to the general public. For example, media imagery and news stories tend to feature apocalyptic or dystopic images of terminators and killer robots, or robots that are humanoid in design, thereby contributing to unrealistic narratives about AI systems and other advanced technologies. A related issue is that such narratives tend to attribute agency to technologies as opposed to humans, with illustrative examples including headlines with questions such as "Can a machine learn morality?". In the worst cases, communicating about AI systems in these ways can create unrealistic expectations and have the consequence of promoting overreliance on AI systems in actual operations. In addition, it risks contributing to already excessive "AI hype" that is characteristic and conducive to techno-solutionist discourses.

From a social justice perspective, it should also be noted that many images used in media stories and marketing materials also stand out in how their narratives and representations of technology relate to representations of "race" and disability, for example. A possible issue here is that images of primarily white robots and humanoid robots equipped with interface elements that are commonly attributed to white persons, for example, can inadvertently reproduce existing, harmful stereotypes concerning the relationship between skin color and intelligence [21]. While this problem is problematic from the perspective of diversity in media representations, it is naturally also closely related to questions of how robots are designed to begin with—such as how they ought to be designed in order to handle questions of "race" with due sensitivity [103].

These problems concern both organizations that market their software products to third parties or that communicate the capabilities of their AI systems to "downstream" operators such as end-users, for example, but also media reporters and organizations that wish to distribute accurate information and critical perspectives concerning emerging technologies. An important point that becomes salient once looking at AI ethics from the lifecycle perspective, however, is that the way AI systems are communicated and marketed, and how the benefits and limitations of AI products and services are communicated to the public, should not be omitted from ethical considerations. Our recommendation is that AI developers and technology companies ought to actively adopt responsible, ethically and socially conscious ways of communicating about AI in these respects. We encourage the reader to examine existing resources, such as the insightful AI Myths website and the Better Images of AI project, which can help responsible agents to communicate about AI systems in a clear and responsible fashion. In addition, a recent blog post by Lakshmi Sivadas and Sabrina Argoub [99] provides media representatives and reporters with concise recommendations on how to report effectively on AI. These include recommendations to build a solid foundation of information on AI and related fields such as technology regulation, avoiding "AI hype" and unreflective adoption of dominant narratives concerning advanced technologies, attributing agency to humans and avoiding the invisibilization of the humans behind AI, and developing a compassionate but critical perspective to covering stories about AI system development and use. On the next page, we will also offer our own concise yet general rules for responsible communication and marketing.

RECOMMENDATIONS:
Three principles for responsible communication

Communication and marketing material should create realistic expectations

Marketing and communication should highlight the unique value of the technology on offer. However, it should not create unrealistic expectations, promote overreliance on the system, or contribute to overall excessive "AI hype" attached to many techno-solutionist discourses (the harms of which have already been highlighted by researchers and journalists). Truthful communication about the affordances and limitations of AI systems serves to promote safe "downstream use" of the system by third parties, enhance trust between different agents, and evince an organization's general commitment to integrity. It is also a precondition for consumers to make meaningful choices regarding system procurement and desirable as a way of mitigating the risk of system misuse—i.e., consumers need to know what a product or service can actually do, and what it is not suitable for. This issue is particularly salient due to the fact that machine learning systems (as most software) are designed for portability across use-contexts and populations. Our recommendation is that, to ensure safe and appropriate procurement and downstream use of AI systems, AI systems and their capabilities—most notably, their limitations—should be described realistically and truthfully. Furthermore, any claims regarding performance and expected impact should be backed up with evidence—preferably with comparisons to existing methods or systems.

Agency should be attributed to humans, not technology

Simply by uttering the words "artificial intelligence", one can often spark lively discussions about sentient robots, doomsday machines, and intelligent artefacts making "moral choices". Unfortunately, while these themes and imageries are grounded in science fiction rather than existing applications or evidence, they seem to still plague many serious discussions about AI as well [117]. These imageries and narratives can create ungrounded expectations about AI but, in the worst cases, also instill fear instead of healthy skepticism. A distinctly problematic trope occurring in these narratives is the attribution of (moral) agency to AI systems instead of the humans who are responsible for them. In particular, anthropomorphic language is often used in this vein—AI systems are described as doing something, for example, rather than being described as instruments that humans have built or which humans use for a given task. While AI systems are more or less functionally autonomous—i.e., they can do certain things without immediate human input—humans are ultimately and thoroughly responsible for them and their effects. In this light, responsible communication about AI should not attribute agency to technologies, respectively.

Language and imagery should promote inclusion and representational diversity

Currently AI systems—including embedded systems such as robots—are predominantly built and portrayed as White in marketing materials and media outputs [21]. Indeed, a quick glance at AI marketing content shows that the majority of the representations of robots include white, male humanoid robots—sometimes weirdly with characteristics of what might be considered stereotypical CEO. This is naturally problematic since in many cases it would simply be quite redundant to build or use the kinds of robots one might encounter in these images—as noted on the AI Myths website, for example, there is a plethora of "images of robots using electronic appliances for which they could have no conceivable need". But, more importantly, from the perspective of social justice, many of the images one might see online neglect the diversity of human beings and cultures by way of primarily portraying Whiteness. Responsible agents should ensure that social values such as multi-culturalism and diversity are recognized and promoted even when communicating about AI. After all, media representations—including also marketing material and communications by the organization—shape conceptions of intelligence, "race", gender, and other categories of identity. In worst cases, the worlds one creates through imagery and discourse can reinforce discriminatory stereotypes, for instance, or marginalize people's experiences. Our recommendation is that marketing material and communications should be designed to promote and celebrate the diversity of human experience and culture, and to avoid reproducing harmful stereotypes and social categorizations.

REFERENCES

[1] Ada Lovelace Institute, AI Now Institute & Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. Retrieved from https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

[2] Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). "A comprehensive review on privacy preserving data mining". SpringerPlus, 4(1): 1-36.

[3] AlgorithmWatch. (2022). Automated Decision-Making Systems and Discrimination. Understanding causes, recognizing cases, supporting those affected: A guidebook for anti-discrimination counseling. Retrieved from https://algorithmwatch.org/en/autocheck/.

[4] Ananny, M., & Crawford, K. (2018). "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". New Media & Society, 20(3): 973-989.

[5] Anderson, Elizabeth, 1999, "What Is the Point of Equality?". Ethics, 109: 287–337.

[6] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J. T., Tsay, J., & Varshney, K. R. (2019). "FactSheets: Increasing trust in AI services through supplier's declarations of conformity". IBM Journal of Research and Development, 63(4/5), 6-1.

[7] Artificial Intelligence Ethics Impact Group. (2019). From Principles to Practice: An interdisciplinary framework to operationalise AI ethics. Retrieved from https://aiethics.site/Library/Prin2Prac.pdf.

[8] Bandy, J. (2021). "Problematic machine behavior: A systematic literature review of algorithm audits". Proceedings of the acm on human-computer interaction, 5(CSCW1), 1-34.

[9] Beauchamp, T. (2019). "The Principle of Beneficence in Applied Ethics". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/.

[10] Bender, E. M., & Friedman, B. (2018). "Data statements for natural language processing: Toward mitigating system bias and enabling better science". Transactions of the Association for Computational Linguistics, 6: 587-604.

[11] Bennett, C. L., & Keyes, O. (2020). "What is the point of fairness? Disability, AI and the complexity of justice". ACM SIGACCESS Accessibility and Computing, (125).

[12] Bergstra, J., & Bengio, Y. (2012). "Random search for hyper-parameter optimization". Journal of machine learning research, 13(2).

[13] Bietti, E. (2020). "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 210-219.

[14] Binns, R. (2018). "Algorithmic Accountability and Public Reason". Philosophy & Technology 31: 543–556. Retrieved from https://doi.org/10.1007/s13347-017-0263-5.

[15] Binns, R. (2020). "On the apparent conflict between individual and group fairness". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 514-524.

[16] Bossert, L., & Hagendorff, T. (2021). "Animals and AI. The role of animals in AI research and application–An overview and ethical evaluation". Technology in Society, 67. Retrieved from https://doi.org/10.1016/j.techsoc.2021.101678.

[17] Brevini, B. (2020). "Black boxes, not green: Mythologizing artificial intelligence and omitting the environment". Big Data & Society, 7(2).

[18] Broussard, M. (2018). Artificial unintelligence: How computers misunderstand the world. MIT Press.

[19] Canca, C. (2020). "Operationalizing AI ethics principles". Communications of the ACM, 63(12):18-21.

[20] Carter, I, (2022). "Positive and Negative Liberty". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/spr2022/entries/liberty-positive-negative/.

[21] Cave, S., & Dihal, K. (2020). "The whiteness of AI". Philosophy & Technology, 33(4): 685-703.

[22] Chouldechova, A. (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". Big data, 5(2): 153-163.

[23] Coeckelbergh, M. (2020). AI ethics. MIT Press.

[24] Coeckelbergh, M. (2021). "AI for climate: Freedom, justice, and other ethical and political challenges". AI and Ethics, 1(1): 67-72.

[25] Cookson, C. (2018, June 9). "Artificial intelligence faces public backlash, warns scientist". Financial Times. Retrieved from https://www.ft.com/content/0b301152-b0f8-11e8-99ca-68cf89602132. Accessed 14.3.2022.

[26] Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. MIT Press.

[27] Crisp, R. (2021). "Well-Being". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/win2021/entries/well-being/.

[28] Darwall, S. L. (1977). "Two kinds of respect". Ethics, 88(1): 36-49.

[29] Digital Regulation Cooperation Forum. (2022). Auditing algorithms: the existing landscape, role of regulators and future outlook. Retrieved from https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.

[30] Dobbe, R. I. J. (2022). "System Safety and Artificial Intelligence". The Oxford Handbook of AI Governance, Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young and Baobao Zhang (eds.). Retrieved from https://doi.org/10.1093/oxfordhb/9780197579329.013.67.

[31] Dryden, J. (N/A). "Autonomy". Internet Encyclopedia of Philosophy, James Fieser & Bradley Dowden (Eds.). Retrieved from https://iep.utm.edu/autonomy/. Accessed 24.1.2023.

[32] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). "Fairness through awareness". Proceedings of the 3rd innovations in theoretical computer science conference: 214-226.

[33] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). "Runaway feedback loops in predictive policing". Proceedings of the 1st Conference Fairness, Accountability and Transparency, PMLR 81: 160-171.

[34] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[35] European Data Protection Board. (2022). Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them. Retrieved from https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-32022-dark-patterns-social-media_en.

[36] European Data Protection Board. (2022). Binding Decision 3/2022 on the dispute submitted by the Irish SA on Meta Platforms Ireland Limited and its Facebook service (Art. 65 GDPR).

[37] Fazelpour, S., & Lipton, Z. C. (2020). "Algorithmic fairness from a non-ideal perspective". Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society: 57-63. Retrieved from https://doi.org/10.1145/3375627.3375828.

[38] Floridi, L. (2019). "Translating principles into practices of digital ethics: five risks of being unethical". Philosophy and Technology, 32(2).

[39] Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). "Privacy-preserving data publishing: A survey of recent developments". ACM Computing Surveys (CSUR), 42(4): 1-53.

[40] Gabriel, I. (2020). "Artificial intelligence, values, and alignment". Minds and machines, 30(3): 411-437. Retrieved from https://doi.org/10.1007/s11023-020-09539-2.

[41] Gabriel, I. (2022). "Toward a Theory of Justice for Artificial Intelligence". Daedalus, 151(2): 218-231.

[42] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). "Datasheets for datasets". arXiv preprint arXiv:1803.09010.

[43] Gebru, T. (2021, December 6th). "For truly ethical AI, its researchers must be independent from big tech". Forbes. Retrieved from https://www.theguardian.com/commentisfree/2021/dec/06/google-silicon-valley-ai-timnit-gebru. Accessed 14.3.2022.

[44] Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books.

[45] Green, B. (2019). ""Good" isn't good enough". Proceedings of the AI for Social Good workshop at NeurIPS, 17.

[46] Green, B. (2021). "The contestation of tech ethics: A sociotechnical approach to technology ethics in practice". Journal of Social Computing, 2(3): 209-225.

[47] Hagendorff, T. (2020). "The ethics of AI ethics: An evaluation of guidelines". Minds and Machines, 30(1): 99-120.

[48] Hardt, M., Price, E., & Srebro, N. (2016). "Equality of opportunity in supervised learning". Advances in Neural Information Processing Systems (NeurIPS), 29.

[49] Haybron, D. (2020). "Happiness". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/sum2020/entries/happiness/.

[50] Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). "The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards". arXiv preprint arXiv:1805.03677.

[51] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). "Improving fairness in machine learning systems: What do industry practitioners need?". Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: 1-16.

[52] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). "Unintended machine learning biases as social barriers for persons with disabilities". ACM SIGACCESS Accessibility and Computing, 125.

[53] Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). "Machine learning and artificial intelligence to aid climate change research and preparedness". Environmental Research Letters, 14(12).

[54] Hutson, J. A., Taft, J. G., Barocas, S., & Levy, K. (2018). "Debiasing desire: Addressing bias & discrimination on intimate platforms". Proceedings of the ACM on Human-Computer Interaction, 2(CSCW): 1-18.

[55] Institute for the Future of Work. (2020). Artificial intelligence in hiring. Assessing impacts on equality. Retrieved from https://www.ifow.org/publications/artificial-intelligence-in-hiring-assessing-impacts-on-equality.

[56] Information Commissioner's Office. (N/A). Guidance on AI and data protection. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/.

[57] Jobin, A., Ienca, M., & Vayena, E. (2019). "The global landscape of AI ethics guidelines". Nature Machine Intelligence, 1(9): 389-399.

[58] Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). "Aligning artificial intelligence with climate change mitigation". Nature Climate Change, 12: 1-10.

[59] Kalluri, P. (2020, July 7th). "Don't ask if artificial intelligence is good or fair, ask how it shifts power". Nature 583: 169. Retrieved from https://doi.org/10.1038/d41586-020-02003-2.

[60] Kaminski, M. E., & Urban, J. M. (2021). "The right to contest AI". Columbia Law Review, 121(7): 1957-2048.

[61] Kenway, J. François, C., Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Bug Bounties for Algorithmic Harms? Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. Retrieved from https://www.ajl.org/bugs.

[62] Kleinberg, J. M., Mullainathan, S., & and Raghavan, M. (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores". Innovations in Theoretical Computer Science Conference (ITCS).

[63] Kousi, T. (2020). "What Can Artificial Intelligence Do for Stray Animals?". Retrieved from https://vetfuturist.com/what-can-artificial-intelligence-do-stray-animals. Accessed 24.1.2023.

[64] Kumagai, J. (2018). "This AI hunts poachers". IEEE Spectrum, 55(1): 54-57.

[65] Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). "Quantifying the carbon emissions of machine learning". arXiv preprint arXiv:1910.09700.

[66] Laitinen, A., & Sahlgren, O. (2021). "AI Systems and Respect for Human Autonomy". Frontiers in Artificial Intelligence, 4:705164. Retrieved from https://doi.org/10.3389/frai.2021.705164.

[67] Lauer, D. (2021). "You cannot have AI ethics without ethics". AI and Ethics, 1(1): 21-25.

[68] Lee, M. S. A., Floridi, L., & Singh, J. (2021). "Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics". AI and Ethics, 1 (4): 529-544. Retrieved from https://doi.org/10.1007/s43681-021-00067-y.

[69] Li, T. C. (forthcoming). "Algorithmic Destruction". SMU Law Review. Retrieved from https://dx.doi.org/10.2139/ssrn.4066845.

[70] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). "Delayed impact of fair machine learning". International Conference on Machine Learning, PMLR: 3150-3158.

[71] Loi, M., & Christen, M. (2020). "Two concepts of group privacy". Philosophy & Technology, 33(2): 207-224.

[72] Luechtefeld, T., Marsh, D., Rowlands, C., & Hartung, T. (2018). "Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility". Toxicological Sciences, 165(1): 198-212.

[73] MacCallum, G. C. Jr., (1967). "Negative and Positive Freedom". Philosophical Review, 76: 312–334.

[74] Mantelero, A. (2018). "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment". Computer Law & Security Review, 34(4): 754-772.

[75] Mantelero, A., & Esposito, M. S. (2021). "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems". Computer Law & Security Review, 41.

[76] Mantelero, A. (2022). Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI. Asser Press: Springer.

[77] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). "A survey on bias and fairness in machine learning". ACM Computing Surveys (CSUR), 54(6): 1-35.

[78] Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). "Algorithmic impact assessments and accountability: The co-construction of impacts". Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency: 735-746.

[79] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). "Model cards for model reporting". Proceedings of the Conference on Fairness, Accountability, and Transparency: 220-229.

[80] Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI". Nature Machine Intelligence, 1(11): 501-507.

[81] Molnar, C. (2020). Interpretable machine learning. Retrieved from https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning-compressed.pdf.

[82] Molnar, P., & Gill, L. (2018). Bots at the gate: A human rights analysis of automated decision-making in Canada's immigration and refugee system. Retrieved from https://citizenlab.ca/2018/09/bots-at-the-gate-human-rights-analysis-automated-decision-making-in-canadas-immigration-refugee-system/.

[83] Munn, L. (2022). "The uselessness of AI ethics". AI and Ethics: 1-9. Retrieved from https://doi.org/10.1007/s43681-022-00209-w.

[84] Müller, V, C. (2021). "Ethics of Artificial Intelligence and Robotics". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

[85] Narayanan, A. (2019). "How to recognize AI snake oil". Arthur Miller Lecture on Science and Ethics.

[86] Nissenbaum, H. (2004). "Privacy as contextual integrity". Washington Law Review, 79(1): 119—158.

[87] North, R. (2017). "Principles as guides: The action-guiding role of justice in politics". The Journal of Politics, 79(1), 75-88.

[88] Nussbaum, M. C. (2006). Frontiers of justice: Disability, nationality, species membership. Belknap Press.

[89] Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence". Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems: 1-13. Retrieved from https://doi.org/10.1145/3313831.3376321.

[90] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". Science, 366(6464): 447-453.

[91] Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). "The fallacy of AI functionality". Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency: 959-972.

[92] Redden, J., Brand, J., Sander, I., & Warne, H. (2022). Automating Public Services: Learning from Cancelled Systems. Retrieved from https://www.carnegieuktrust.org.uk/publications/automating-public-services-learning-from-cancelled-systems/.

[93] Regan, T. (2004). The case for animal rights. University of California Press.

[94] Riley, S. (N/A). "Human dignity". Internet Encyclopedia of Philosophy, James Fieser & Bradley Dowden (Eds.). Retrieved from https://iep.utm.edu/human-dignity/. Accessed 24.1.2023.

[95] Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Lucconi, A. S., Maharaj, T., Sherwing, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt., J. C., Creutzig, F., Chayes, J., & Bengio, Y. (2022). "Tackling climate change with machine learning". ACM Computing Surveys (CSUR), 55(2): 1-96.

[96] Ryan, M., & Stahl, B. C. (2020). "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications". Journal of Information, Communication and Ethics in Society, 19(1): 81-86. Retrieved from https://doi.org/10.1108/JICES-12-2019-0138.

[97] Scanlon, T. M. (2018). Why does inequality matter?. Oxford University Press.

[98] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). "Fairness and abstraction in sociotechnical systems". Proceedings of the Conference on Fairness, Accountability, and Transparency: 59-68.

[99] Sivadas, L., & Argoub, S. (2021, May 5th). "How to report effectively on artificial intelligence". Retrieved from https://blogs.lse.ac.uk/polis/2021/05/05/how-to-report-effectively-on-artificial-intelligence/. Accessed 24.1.2023.

[100] Sloane, M. (2021, March 17th). "The Algorithmic Auditing Trap. Medium". Retrieved from https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d. Accessed 24.1.2023.

[101] Smith, P., & Smith, L. (2021). "Artificial intelligence and disability: too much promise, yet too little substance?". AI and Ethics, 1(1): 81-86.

[102] Sokol, K., & Flach, P. (2020). "Explainability fact sheets: a framework for systematic assessment of explainable approaches". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 56-67.

[103] Sparrow, R. (2020). "Robotics has a race problem". Science, Technology, & Human Values, 45(3): 538-560.

[104] Stoljar, N. (2018). "Feminist Perspectives on Autonomy". The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/win2018/entries/feminism-autonomy/.

[105] Sætra, H. S. (2021). "AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system". Sustainability, 13(4).

[106] Thomas, R. L., & Uminsky, D. (2022). "Reliance on metrics is a fundamental challenge for AI". Patterns, 3(5). Retrieved from https://doi.org/10.1016/j.patter.2022.100476.

[107] Vallor, S., Green, B., & Raicu, I. (2018). Ethics in Technology Practice. The Markkula Center for Applied Ethics at Santa Clara University. Retrieved from https://www.scu.edu/ethics-in-technology-practice/. Accessed 9.2.2022.

[108] Van Noorden, R. (2020). "The ethical questions that haunt facial-recognition research". Nature, 587(7834): 354-359.

[109] van Wynsberghe, A. (2021). "Sustainable AI: AI for sustainability and the sustainability of AI". AI and Ethics, 1(3): 213-218.

[110] Veale, M., Binns, R., & Edwards, L. (2018). "Algorithms that remember: model inversion attacks and data protection law". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133).

[111] Verma, S., & Rubin, J. (2018). "Fairness definitions explained". 2018 IEEE/ACM International Workshop on Software Fairness (FairWare): 1-7.

[112] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). "The role of artificial intelligence in achieving the Sustainable Development Goals". Nature Communications, 11(1): 1-10.

[113] Wachter, S. (2020). "Affinity profiling and discrimination by association in online behavioral advertising". Berkeley Technology Law Journal, 35(2). Retrieved from https://dx.doi.org/10.2139/ssrn.3388639.

[114] Wachter, S., Mittelstadt, B., & Russell, C. (2021). "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI". Computer Law & Security Review, 41. Retrieved from https://dx.doi.org/10.2139/ssrn.3547922.

[115] Wachter, S., Mittelstadt, B., & Russell, C. (2020). "Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law". West Virginia Law Review, 123(3). Retrieved from https://dx.doi.org/10.2139/ssrn.3792772.

[116] Wagner, I., & Eckhoff, D. (2018). "Technical privacy metrics: a systematic survey". ACM Computing Surveys (CSUR), 51(3): 1-38.

[117] Wallenborn, J. T. (2022, May 17th). "AI as a flying blue brain? How metaphors influence our visions about AI". Digital Society Blog. Retrieved from https://www.hiig.de/en/ai-metaphors/. Accessed 12.12.2022.

[118] Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Ringel Morris, M., Rankin, J., Rogers, E., Salas, M., & West, S. M. (2019). Disability, bias, and AI. AI Now Institute. Retrieved from https://ainowinstitute.org/disabilitybiasai-2019.pdf.

[119] Wieringa, M. (2020). "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 1-18.

[120] Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). "Building and auditing fair algorithms: A case study in candidate screening". Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency: 666-677.

[121] World Economic Forum. (2020). Ethics by Design: An organizational approach to responsible use of technology. [White paper]. Retrieved from https://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf.

[122] Zeng, Y., Lu, E., & Huangfu, C. (2018). "Linking artificial intelligence principles". arXiv preprint arXiv:1812.04814.