Anthropomorphization in the context of human cooperation with intelligent machines

Bachelor’s Thesis at the
Faculty of Philosophy, Philosophy of Science
and the Study of Religion

Ludwig-Maximilian University Munich

Benjamin Schiffer

Munich, 12 January 2020

Supervisor Dr Jurgis Karpus

Contents

Chapter 1

Introduction

1.1 A motivation for human cooperation with machines

Intelligent machines are on the rise. Ever since the first computers were invented, humans have endeavoured to make algorithms more intelligent. Artificial intelligence has been a central goal of computer science for many decades and today we are closer to this goal than ever before. We are already using artificial narrow intelligence (ANI) constantly when browsing the web or talking to intelligent home assistants such as Amazon's Alexa. The way humans interact with intelligent machines has therefore never been as important as today.[1]
Rahwan et al. (2019) have argued recently for the creation of the new scientific field of *machine behaviour* as studying machine behaviour seems critical to maximizing the potential benefits of artificial intelligence (AI) in our society. In order to do so, an interdisciplinary approach is necessary. The cooperation of humans and intelligent machines is a form of hybrid human-machine behaviour that plays a big role within the field of machine behaviour.

In this work we will attempt to provide a philosophical perspective on human-machine cooperation by looking at how human behaviour is shaped by machines and how humans can be nudged towards cooperating with machines in a better way.
Already today and especially in the future, we can only harness the benefits of AI if we care about what the cooperation between humans and machines will look like. The challenges humanity is facing (e.g. Lenton et al., 2019) are so pressing that we, both individually and as a society, need to address how we imagine human-machine cooperation in the 21st century. In this sense, we adopt a consequentialist perspective on the use of modern AI technology.

[1] Whenever we will refer to machines, we mean intelligent machines that possess at least artificial narrow intelligence. If not clear otherwise from the context, a robot here is an intelligent machine which has some degree of autonomy, and a computer program or algorithm is understood to usually involve state-of-the-art ANI

1.2 The argument for anthropomorphizing machines

We will use an inductive approach and start from general issues involved in artificial intelligence today. We are convinced that this approach is well-suited to provide the necessary amount of background information and then discuss the use of anthropomorphization of machines to improve human-machine cooperation. We briefly present the main points of the following chapters here:

- We start with the history of artificial intelligence to show how rapid the development of intelligent machines is. We discuss the ubiquitous and opaque nature of intelligent machines. Finally, we argue why human-machine cooperation is crucial and that we need to address this cooperation from the perspective of the human agent (premise 1).

- After narrowing down our focus to human-machine cooperation, we discuss issues around it. Cooperation with machines in many cases is less efficient than it could be. We discuss empirical evidence for algorithm aversion and also show that efficient cooperation is possible (premise 2). We raise the point that humans cooperate more eagerly when dealing with a machine which they anthropomorphize.

- We therefore turn towards the field of anthropomorphization and discuss the strong human tendency to humanize machines (premise 3). We present a selection of anecdotal and empirical evidence regarding personification and perceived agency of machines as well as empathy and abuse towards them.

- We synergize above points and discuss positive and negative examples of the anthropomorphization of machines. We conclude that for ethical reasons, the use of humanoid features in machines needs to be decided on a use-case basis (conclusion). Understanding anthropomorphization as a form of nudging, we build an argument that for the use-case differentiation to be successful, humans need to be able to selectively anthropomorphize only humanoid machines (question concerning the conclusion).

- We finally present a study proposal to test this ability of humans to distinguish machines into different categories depending on whether they are endowed with humanoid features (suggestion regarding above question).

This philosophical work touches upon many fields, among these computer science, neuroscience and psychology. We hope to have chosen the right balance and depth to make our argument as clear as possible without wandering off too much into the fascinating subfields. We also wish that this work can contribute to making individuals and society benefit from the advances in intelligent machines in the best possible way.

Chapter 2

The rise of intelligent machines

> "Machines take me by surprise with great frequency"
>
> Alan Turing, 1950

In this chapter, we are going to argue why human-machine cooperation is important and why it will continue to be so in the foreseeable future. We will outline in brief the development of intelligent machines to show the pace at which the field advances. We also discuss the ubiquity of intelligent machines in our society today.

Another important point will be the opaque nature of intelligent machines as it is much harder or even impossible to understand the inner workings of modern machine learning algorithms.

2.1 A brief history of intelligent machines

2.1.1 GOFAI and the quest for artificial intelligence

While the dream of building an artificially intelligent machine has been around for much longer[1], if one was to put a beginning to artificial intelligence, Alan Turing's famous 1950 paper seems to be a sensible choice (Turing, 1950). In his manifesto for AI, he not only proposed the *The Imitation Game* which is also referred to as the Turing test[2], but also identified the main questions regarding intelligent information processing (Boden, 2018, p. 7f.).

The term *artificial intelligence* was first used at the Dartmouth Conference in the summer of 1956. In Dartmouth, John McCarthy, Marvin Minsky and other researchers defined the birth of AI and went on to become influential thinkers in the field. It emerged what is now referred to as good old-fashioned artificial intelligence (GOFAI). The approaches in these

[1] See for example of the mechanical turk. This 18th century machine had in fact a human inside but was able to fool many into believing that it was an intelligent machine (The Editors of Encyclopaedia Britannica, n.d.).

[2] The Turing test is a concept to assess the intelligent behaviour of a machine. Very briefly put, if a human after a short natural language conversation with a machine is not able to tell whether he was chatting with a machine or a human, then the machine passes this intelligence test.

early AI programmes were based on common sense and reasoning and the programs achieved impressive feats at complex geometry and algebra tasks (Levesque, 2018, p. 4f.). Linguistics played a huge role as computers were supposed to communicate in natural language. A famous milestone was the ELIZA chatbot which was able to fool people into communicating with a computer psychotherapist even though the chatbot was not yet very sophisticated. This virtual Rogerian[3] psychotherapist would ask simple questions and people engaged eagerly in conversations with the chatbot, telling it intimate secrets (Weizenbaum, 1966; Weizenbaum, 1976, p. 3f.).

Without going into the details of the early years of AI research, it is important to note that - and this is something we still see today - the first successes of computers were perceived as something astonishing, extraordinary and also seemingly intelligent (Russell & Norvig, 2016, p. 18). While it is easy to merely consider the achievements of GOFAI from a modern perspective as still being very unintelligent, this is misleading. It is true that GOFAI did not create powerful machine intelligence and there were a number of setbacks as the great optimism of the first years did not lead to general artificial intelligence as soon as had been hoped. However, we would argue that this follows a pattern. A task might initially seem to require some sort of intelligence to be performed. Once an algorithm is capable of the same achievement, the general view held of this task changes and the task is merely seen as something mechanical where in fact no real intelligence was needed in the first place.

This might be interpreted as an ongoing underestimation how hard it is to create artificial intelligence. While this is without doubt true, it also implies that we constantly reevaluate our self-image as humans. This view is also held by Kaplan (2004) who agrees that "progress in artificial intelligence may significantly change what we thought were features unique to humans." (p. 478) The less unique human intelligent features are, the more likely it becomes that it is feasible to create an artificial intelligence that rivals the intelligence of humans. The remainder of this chapter and the rapid progress in building intelligent machines, should be seen from this perspective.

2.1.2 Artificial neural networks and the second AI revolution

An important technical development in the field of AI are artificial neural networks (ANN). The advances of the last 25 years in artificial intelligence research are strongly connected with ANN which work very differently from algorithms in the GOFAI era (cf. Boden, 2018, p. 69).

The idea of ANN goes back to the first description of a computational neuron model, the McCulloch-Pitts-neuron (McCulloch & Pitts, 1943) where Bertrand Russel's propositional logic and Charles Sherrington's theory of neural synapses were united with Alan Turing's work. Logical values of true or false were mapped onto the on/off activity of brain cells as well as the zero/one state in Turing machines. Together with Sherrington's belief that neurons have fixed thresholds, the many connected computational neurons manifest a neural

[3] Person-centered approach in psychotherapy named after Carl Rogers.

net which offers computational use (Boden, 2018, p. 8f.).
Neural networks are able to learn through adaptive changes in the weights of the individual neural connections and rarely even in the connections themselves. This learning process usually works in a way that goes back to the neuropsychologist Donald Hebb who developed the concept that neural connections that are used more often are being strengthened.
The recent impressive advances in machine learning are attributed to a big part to *deep learning* which was initiated by Jürgen Schmidhuber in the 1980s. The term deep learning refers to the fact that these networks use a multi-layered approach. Between the input and the output layers lie a number of hidden layers. When the number of hidden middle layers in a learning network is large, it is called *deep learning*. Such a network when trained on data exhibits a very complex inner structure (cf. Boden, 2018, p. 74ff.).

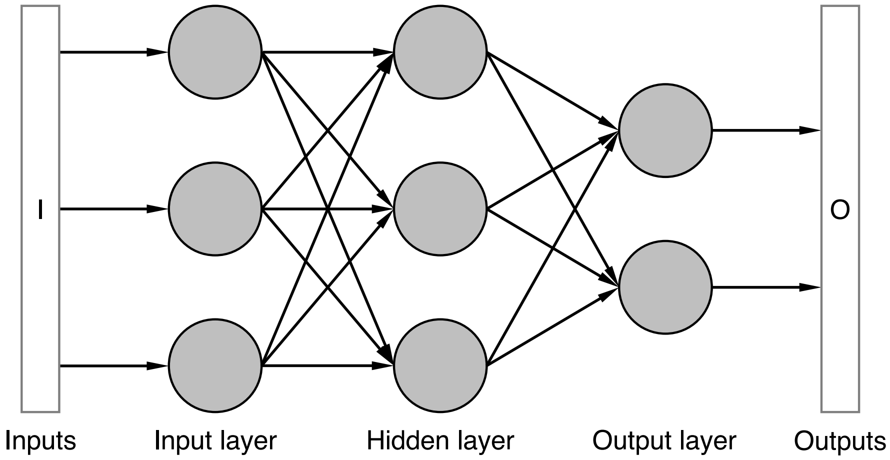

Figure 2.1 – Diagram representing a neural network with an input layer, one hidden layer and an output layer. The nodes of the network are connected by computational neurons. Deep learning uses neural networks with many hidden layers.

While the explanations given above are more theoretical, we shall offer a more intuitive explanation of how deep learning works. So, let us imagine that we are trying to build a computer program that is very good at recognizing cats in a given collection of pictures. The classical (non-machine learning) approach would be to tell the program what to look for when searching for a cat. We might achieve this by thinking of two characteristic angles of the cat, the face and the body shape from the side. Now in order to recognize the face, we will further assume that a cat face features greenish eyes, vertical pupils, a triangular shaped nose, and so on. Whenever it recognizes the right spatial combination of those features, the program should deduce that it is seeing a cat. Similarly, we could come up with a recognition algorithm for the body shape and then when combining these two characteristic angles of a cat, we would hope to detect the cats in the pictures with very high probability (Levesque, 2018, p. 2f.).

In machine learning, however, the approach is quite different. Instead of manually defining what a cat looks like, a deep neural network is supposed to deduce these features on its own after being trained on a huge data set. This does not imply logical reasoning but merely a statistical process using a vast number of connection computational neurons. With each iteration of the learning algorithm, the neural network sees another picture of the training data set and takes a guess based on what it has learned so far (i.e. its connections and weights). Then after comparing the guess with the true result, whether this was a cat or not, the weights of the neural network are updated. In practice this learning process might not always work as smoothly as described here, but in theory the algorithm converges to minimal loss. The loss describes the gap between perfect detection and actual detection. This approach of feeding a suitable neural network with large amounts of well-sorted data has turned out to work surprisingly well (cf. Spiegelhalter, 2019, Ch. 6).

Interestingly, the layered pattern detection seems also to be similar to the mechanism of how the brain works. Especially human vision is fairly well understood and, very broadly speaking, the process of human vision is as follows:

When we see an object, say the word "APPLE", it is analysed by human vision in a multi-layered approach. Individual small patterns are interpreted first, this could be the 45° stroke within the letter "A". Then, after also having interpreted the other two lines which together compose the letter "A", these strokes are being combined in order to recognize the pattern of each individual letter. Finally, all the letters combined will form the whole word APPLE which is then recognized by the brain (Kurzweil, 2012, p. 41f.).

This close relation between advances in artificial intelligence and neurophysiology is another clue that AI might working in the right direction and that the exchange between AI and brain researchers seems promising for the quest for AI (see also Savage, 2019).

The modern revolution in machine learning was made possible also due to the increased availability of massive data collections, new techniques to handle big data and raw computing power. In the same way the successes of GOFAI were impressive and seemingly intelligent, the advances of modern AI underline how we shifted our view in what is intelligent behaviour and what computers can achieve.

There have been a number of milestones in the last few years where deep learning played a big role. IBM's product *Watson* is used in call centres and as an assistant in medical applications. It gained a lot of popular attention when it was able to beat the best human players at *Jeopardy!*[4] in 2011.

Deep Learning is also used by many Google services, especially at Google DeepMind. Their program AlphaGo beat the world champion in the strategic board game Go in 2016 (Boden, 2018, p. 57f., 79f.). Especially one move that AlphaGo made, move 37, was widely debated. It was the crucial move for AlphaGo to win and to most human commentators this move seemed highly unconventional if not an error. It turned out that the artificial intelligence

[4] Jeopardy! is an American TV quiz show where the players are given an answer and have to guess the relevant question. This scenario is opposite to usual quizzes and much harder for computers to solve.

had in fact made a very smart decision and was able to win the game. The originality of the move was often referred to as something inhuman, out-of-this-world or transcendental (Metz, 2017).

A recent example of AI is the success at the card game poker, where a program called Pluribus significantly outperformed human players in a six player poker game (Brown & Sandholm, 2019).

2.2 The ubiquity and the opacity of intelligent machines

2.2.1 Intelligent machines are everywhere

Besides these outstanding AI milestones and even more importantly, algorithms are constantly present in our daily lives. We are exposed to them in many online applications, e.g. dynamic pricing in web shops and personalised news rankings. We see how social media bots alter the way information is perceived and AI determines which profiles are shown to you in online dating. Car manufacturers around the world are in an arm's race to build level 4 and 5 self-driving cars[5] which rely heavily on modern AI technology (cf. Rahwan et al., 2019). Meanwhile, intelligent assistants conquer our private spaces with Amazon having sold over 100 million Alexa devices already. If we include the assistants installed on Android phones and iPhones, it is estimated that the total number of intelligent assistants will surpass the world population before 2023 (Perez, 2019).

Another important fact regarding the ubiquity of intelligent machines is that the entry level barrier for individuals to get started with deep learning or modern AI technology in general could not be lower. It takes only 30 minutes and a computer connected to the internet to install a developing environment for the very popular programming language Python and to download a machine learning framework. There is a vast number of helpful frameworks available with the most well-known being Tensorflow and Pytorch. Tensorflow was presented by Google Brain, the deep learning team at Google (Abadi et al., 2016) and is used a lot in Google products, while Pytorch is a framework developed by Facebook (Paszke et al., 2019). In recent years, it has become very easy for anybody around the world to run their first deep learning algorithm, e.g., an algorithm for the classification of images of clothing, in virtually no time (see Google, n.d.).

This should not at all imply that deep learning is easy, and building robust and useful deep learning algorithms can be a big challenge. However, the creators of the machine learning frameworks have paved the way for an increased usage of deep learning in a large number of applications.[6]

[5] The Society of Automobile Engineers (SAE) defines five levels of driving automation. Level 4 and 5 correspond to high and full automation, respectively, and are the highest automation levels. (SAE International, n.d.)

[6] It might seem like a valid objection that the number of people who are able to write a deep learning algorithm is still relatively small. But then this does not change the fact that due to the easiness of implementing such an algorithm, already modern AI technology affects virtually all human beings in most regions worldwide on a regular basis.

A third point why AI technology will continue to be everywhere is that algorithms diffuse and scale very easily. New AI algorithms are usually reproduced within days or weeks after being published and there is a very open culture within the field of AI research so that many papers publish their source codes. Besides, an AI system is typically efficient and scalable. That means that, e.g., a typical facial recognition system can be applied to many different camera feeds for much less than the cost of hiring a human analysist once the algorithm has been developed and trained (Brundage et al., 2018).

2.2.2 The opaque nature of intelligent machines

Now we will turn towards the different nature of intelligent machines due to their deep learning features. In such an unsupervised learning setup, we would not know what led the machine to detect the cat in a picture. This is because it is not possible to just match parts of a neural network with specific properties of the cat, such as the tail. Instead, deep learning algorithms are used as a black box where we have little to no clue of what is happening inside.

Using algorithms as a black box is nothing new. In his 1976 book "Computer Power and Human Reason" Weizenbaum warns against the use of algorithms especially in the military context. He mentions the case of the U.S. military using a computer program in the Vietnam war to "declare free-fire zones, that is, large geographical areas in which pilots had the 'right' to kill every living thing." The operators of these computer programs did not understand what went on within the computers and used them as a black box. Yet, they entrusted this black box with life-or-death decisions (Weizenbaum, 1976, p. 238).

This is an example of the unethical use of technology because the accountability for military actions that follow such computers advice is unclear. Theoretically, however, it was very well possible to have an expert, say the creator of the program, analysing the inner workings of this U.S. military algorithm. It became a black box only by the specific set-up it was used in, it became a black box by the ignorance of the operators.

Today's deep learning algorithms generally no longer allow this possibility. Instead, the very nature of deep learning leads to it being a black box. There are advances in the field of explainable AI (XAI) (cf. Samek, Wiegand, and Müller, 2017) which seek to address this issue and make the statistical reasoning of AI understandable for humans. Yet, the opacity of the inner workings of deep learning algorithms is, at least for the time being, inherently connected to these intelligent machines.

2.2.3 Algorithmic bias as a risk for cooperation

We do know that machines have become very intelligent, yet, there is one massive caveat. Machine learning algorithms have repeatedly been shown to reproduce biases that are introduced mostly by the data they are trained on.[7] Due to the discussed opaqueness of deep

[7] For a discussion of different existing biases and approaches at how to overcome them, see Danks and London (2017).

learning tools, there is the risk of a dangerous feedback loop where biased algorithmic advice reinforces unacceptable discrimination in society.

Regarding gender bias and discrimination, we see, for example, how voice assistants such as Siri or Alexa have female voices by default while IBM's Watson had a male voice. Generally speaking, the over-representation of men in the design of technology could present great harm for the advances in gender equality. This is because the developers of these algorithms are mostly unaware of research in how gender ideology is embedded in language as Leavy (2018) argues. In section 2.2 we discussed how easily AI systems diffuse, which can lead to discriminating algorithms being distributed around the world very quickly. To prevent diminishing trust in AI systems, these algorithms need to consistently produce bias-free advice. Fair representation of women and minorities in the development of technology is therefore crucial.

To overcome the black box that deep learning usually is, a growing number of researchers are trying to build explainable AI (XAI). The idea here is that when we are told specific reasons for the decision that a machine makes, we should be able to reconstruct the reasoning. In the example of the cat image detection algorithm, this would imply, that if we knew that the algorithm had falsely deduced that cats were always grey, due to bias in the training data set, we might be able to correct this. Hopefully, this would make it much easier to combat discriminating algorithms.[8]

Algorithmic bias is a risk for human-machine cooperation for two reasons. First, the performance of intelligent machines is not as good as it could be and, and second, human trust in machines will be damaged due to the biases present in the intelligent machines. Algorithmic bias will not be a focus of this work but these considerations are important if we want to use AI as a useful and smart tool.

2.3 The need for human-machine cooperation

We have investigated the past and current state of intelligent machines and found that there is an impressive dynamic towards ever more intelligent systems. Already today intelligent machines are ubiquitous in our everyday private and business lives. With a technology so powerful, it is necessary to ask how we intend to use it. We will argue here that a crucial aspect which we will focus on is human-machine cooperation.

Machines have already been able to produce forecasts superior to human ones for a while by using standard statistical methods (Grove, Zald, Lebow, Snitz, & Nelson, 2000). We discussed that today AI players are able to outperform humans in the complex games of

[8] We would like to point towards two interesting points. There is recent work on inferred causality in machine learning. As our societal norms and laws are not based on stochastic reasoning but on causal considerations, this seems to be a promising direction for advancing fair algorithms (Kilbertus et al., 2017).
Another reason why XAI might become more relevant is the EU General Data Protection Regulation (GDPR) of 2018. The GDPR gives all residents of the EU the right to receive an explanation of how an automated decision was made. What this means exactly for opaque deep learning algorithms has not been decided in court yet. However, the GDPR definitely helps the push for XAI (Wagner, n.d.).

Go and Poker and also, generally, machines have been closing the gap to human skills in recent years. The fields of image recognition, speech recognition, abstract strategy games, a large number of real-time video games (such as Atari games) are further examples where algorithm performance now exceeds human performance.

At the same time, there are a number of fields where the direct comparison human vs. machine is difficult, where there is only slow progress in AI or where no systematic comparison and trend analysis has been made yet. For instance in the fields of visual question answering and translations, humans still come out top (Peter Eckersley et al., 2017).

This indicates that, at least for the near future, humans and machine have expert skills in different fields. Hybrid human-machine teams can combine the different strengths. Therefore, the strong performance of intelligent machines is an argument for an increased cooperation between machines and humans.

An insightful example regarding the roles of machines in our future society comes from chess. In 1997, chess programs had advanced so much that IBM's chess computer DeepBlue was able to defeat the then chess world champion Garry Kasparov. Following the advancement of chess machines to superhuman chess skills, machines where increasingly used as sparring partner for humans. Through training with very smart algorithms, humans were able to deepen their knowledge and become better in chess than they would have been without chess machines (cf. Levesque, 2018, p. 131f.).

This could be a strong analogy to human-machine cooperation in general. While the human goal in chess is simply excelling at this strategic board game, in general we strive for individual and collective happiness, safety and prosperity. We want to reliably rule out safety-critical defects in manufacturing processes, detect early-phase cancer as soon as possible, eliminate road toll through automated driving and accelerate technological advances in renewable energies to help make the planet sustainable as quickly as possible. Then, just like in chess, it is reasonable and from a consequentalist perspective even required to combine the different strengths of machines and human as the cooperation between humans and machines is able to lead to better results than human-human cooperation alone. It is our aim to help providing the right framework for human-machine cooperation so that the outcome of this cooperation will be as favourable as possible.

We will address one possible counter argument against human-machine cooperation here. We argued that the current different expert skills of machines and humans differ and therefore cooperation is needed. If artificial intelligence might very soon have superhuman skills in all areas and evolve into artificial general intelligence (AGI), the goal of cooperating with humans might no longer be necessary. Humans would be made superfluous for many tasks as they would be too inefficient.

There are a number of technological visionaries that warn against such a technological singularity when machine intelligence surpasses human intelligence. From this point onwards, AI would be able continue to develop itself at ever increasing rates. Such artificial superintelligence (ASI) would then be extremely powerful which might constitute a significant threat to humanity.[9] We will not focus on AGI or even ASI here because of two reasons. First, it is not certain that such scenarios will actually arise, and if they do, this will very likely occur only in a few decades from now.[10] We know from the history of artificial intelligence that creating AI always turned out to be more difficult than expected. Therefore, it seems reasonable to treat the above estimates with great care. Secondly, already today we see many examples of ANI leading to real challenges which need be addressed. Hence, we should focus on human-machine cooperation now instead of hypothetical scenarios.

[9] Nick Bostrom uses the example of a super-intelligent machine with the task of producing an exact number of paperclips. To make sure that it did not miscount the already produced paperclips, the machine will produce spare ones. However, as the probability to miscount never drops to zero, the machine will need to produce an infinite number of paperclips thereby using up all resources of the earth. While this likely would not match the layman's definition of super-intelligent behaviour, these kinds of threats might actually become relevant in the (far) future (Bostrom, 2016, p. 150ff.).

[10] Expert opinions on the possible timescale of such events vary greatly. Across different studies around 10% of the experts in AI predict human-level machine intelligence around the years 2020-2024, 50% around 2040-2050 and 90% around 2065-2093 (Müller & Bostrom, 2016).

Chapter 3

Human acceptance of or reluctance against algorithms

> "Never trust anything that can think for itself if you can't see where it keeps its brain"
>
> Arthur Weasley in J.K. Rowling: Harry Potter and the Prisoner of Azkaban, 1999

We have already argued in the previous chapter why it is so important that humans and machine cooperate with each other. However, there are several problems regarding the cooperation with intelligent machines. For example, the issue of opaqueness of AI, which refers to the black box nature of deep learning algorithms, already came up. Besides these issues that lie more on the technical side, a very fundamental aspect of human-machine cooperation is the inclination of the human agent for cooperation.

As we will see in this chapter, in fact humans show unwillingness to cooperate with machines in some scenarios. We have already discussed that machines are superior to humans at certain tasks. Then, such reluctance can be costly if cooperation would have led to better outcomes and is accordingly a manifestation of irrational human behaviour.

Here, we are going to investigate the possible obstacles on the human side for human-machine cooperation with the goal to overcome them.

3.1 Do humans show algorithm aversion?

There are many different forms of cooperation imaginable. A very simple form of cooperation between two agents is the action of giving and receiving advice. This is a rather unidirectional way of cooperation, yet, it is very frequent and especially important when considering cooperation between a human and a machine. This form of cooperation is between two agents where one agent possesses knowledge and the other agent needs to assess and tailor this knowledge and output to a real-world problem. Because of superior machine computing skills, some machine advice such as forecasts have been superior to human forecasts already for decades. With machine forecasts becoming better every year, intelligent machines will exceed human skills in ever more areas. This will make the question of how humans deal with machine advice even more important in the future.

The earliest discussion of human distrust in algorithmic advice most likely goes back to the psychologist Paul Meehl. In 1954, he published a study on what he referred to as the conflict between clinical and statistical prediction (Meehl, 1954). He "had analyzed whether clinical predictions based on the subjective impressions of trained professionals were more accurate than statistical predictions made by combining a few scores or ratings according to a rule" as Daniel Kahneman explains (Kahneman, 2011, p. 222). Meehl found that even merely a linear statistical model was able to outperform human judgement. Clinical psychologists reacted with shock, hostility and disbelief to Meehl's findings.

It took many decades until mistrust in algorithms was investigated empirically.[1] We present an overview of the most relevant recent findings here.

It has been shown that most people do not choose a statistical model for making a forecast unless they have more confidence in the model's forecast than in the human forecast (Dietvorst, Simmons, & Massey, 2015). Participants that were less convinced of the algorithm grew more reluctant after they witnessed the model perform than participants that had great confidence in the model from the outset. This reluctance was coined by Dietvorst et al. (2015) as *algorithm aversion*.

However, algorithm aversion does not imply a general aversion to using algorithms at all. In a study, Logg et al. (2019) found that in fact people were inclined to use algorithmic advice. Interestingly however, participants who were experts in their field showed different behaviour. They were reluctant to use algorithms which resulted in the expert's predictions being worse than the predictions of laypeople who had made use of algorithmic advice.

In another experiment Prahl and Van Swol (2017) confirmed that people do accept algorithmic advice, but they found evidence that automation trust is, in fact, heavily discounted after seeing a machine err. This is remarkable due to its irrationality. Prahl and Van Swol (2017) used advice response theory (ART) as a framework to think about human-machine cooperation. ART was originally designed as a framework to investigate inter-human trust. They argue that automation trust is an important factor in the question of algorithm aversion. To explain the discounting of automation trust, they refer to the "perfections schema" by Madhavan and Wiegmann (2007). This pattern suggests that human forecasters have very high expectations regarding algorithmic advice and assume it to be perfect. Humans, however, are known to be fallible and therefore met with to lower expectations. Then, for the human forecasters, an error generated by the algorithm is an unexpected error which was not foreseen as would have a human error. As a result, trust in the algorithm is shattered and the human forecasters become reluctant to use algorithmic advice even though it is still superior to human advice.

It should be added here that Dietvorst, Simmons, and Massey (2016) made an interesting

[1] For a more complete overview of relevant studies conducted over the last 20 years see Logg, Minson, and Moore (2019).

finding when looking at how human acceptance of algorithmic advice might be fostered. They found that people are more eager to use even imperfect algorithms as long as they have the option to slightly modify these algorithms. Interestingly, the extent by which they are able to modify does not change how frequently the algorithm is used. It seems to be the feeling of behaving in an active instead of a passive way which is important to humans. Besides, when humans were able to modify the workings of the algorithms in the Dietvorst et al. (2016) study, they reported higher satisfaction with the cooperation.

The presented research in the field of algorithm aversion gives useful insights on human-machine cooperation. Algorithm aversion points to the fact that humans consistently behave irrationally by not always accepting superior machine advice. It also includes the finding how humans discount automation trust after seeing a machine err. In this context it is troublesome that especially experts express algorithm aversion and overestimate their own skills. They are the ones who should be trained to know better as their behaviour might have the most severe consequences, e.g. in the screening of tumor cells in radiology.

3.2 Human-machine cooperation in strategic games

We have looked at the act of receiving advice as a form human-machine cooperation already and found evidence for algorithm aversion. However, advice is not the only way humans co-operate with machines and we will need to look at different settings as well. We are interested in more strategic forms of cooperation between two agents. Testing human-machine cooperation empirically always requires a certain abstraction from a real-life scenario. Strategic games mimic strategic cooperation and make it rather easy to measure performance. Here, we will look at evidence that modern algorithms are able to achieve superhuman performance in these strategic games.

In the previous chapter we mentioned AI milestones that include the defeat of a human contestant in strategic games such as Chess, Go or Poker. These games are zero-sum encounters (cf. Crandall et al., 2018). For many other games, however, cooperation between the players is needed to perform well. Such games can provide a laboratory setting for the investigation of human-machine cooperation.

Indeed, Crandall et al. (2018) found that their intelligent algorithm (called S#) was able to form a cooperative relationship with a human player. The machine was able to interact with the human via non-binding costless signals, also referred to as cheap talk.

Crandall et al. (2018) used a simple stochastic game called *Block Game* in their study. The goal of this game was for each player to achieve the highest possible point value for their set of blocks. Similary to the more simple well-known prisoner's dilemma, in Block Game many cooperation scenarios are possible. These scenarios vary in the fairness of the players and in the efficiency of the outcome. Especially, unfair game-play is possible to prevent the other player from getting more points, and beneficial outcomes for both players are also possible.

The game was played over several rounds and the development of the average payoff was compared between the two set-ups human vs. human and S#-[2] vs. the human. Notably, S#- successfully uses cheap talk to consistently build cooperative relationships with the human player. The average payoff of the set-up S#- vs. human was higher than in the inter-human gameplay.

The authors looked into the reasons for S#'s remarkable performance. They found that S# mostly stuck to cooperative gameplay after cooperation had been established in several rounds in contrast to human players who tended to defect more often from cooperation. Also, S# was committed to what it communicated via cheap talk while a sizable portion of the human participants did not.

This research indicates that beneficial cooperation between humans and intelligent machines is possible. Similar to the case of giving advice, the human agent seems to be mainly responsible for inefficient cooperation.

We saw that algorithms outperform humans even at the task of cooperating with a human being in strategic games. This is a valuable insight as it underlines our argument for human-machine cooperation. Merging the concept of algorithm aversion with performance in strategic games, we should be wondering about human reluctance towards algorithms in such a cooperation game.

Ishowo-Oloko et al. (2019) showed that people tend to cooperate better when they were left in the dark regarding the true identity of their cooperation partner. They conducted a canonically iterated prisoner's dilemma where participants played against a machine[3] or a human. The true nature of their interaction partner was disclosed to one half of the participants only while the other half was given inaccurate information.

They found that a machine passing as a human was more efficient than a real human (for similar reasons as in the experiment by Crandall et al. (2018)), however, only as long as the participant was misinformed about the machine's true nature. With transparency on the actual machine nature, the efficiency drops markedly leading to worse cooperation rates than for a human. The magnitude of the effect was about 10 percentage points.

As a matter of fact, we see that human algorithm aversion prevents the best outcome of human-machine cooperation. This raises the ethical question of a transparency-efficiency trade-off.

3.3 Overcoming algorithm aversion

We have seen how humans are repeatedly responsible for inefficient human-machine cooperation and that the human reluctance to engage with algorithms is irrational. We are unaware of empirical studies investigating whether algorithm aversion fades when humans are exposed

[2] S#- was an earlier version of their S# algorithm which was only able to generate cheap talk but was not able to respond to it.

[3] The actions of the machine were calculated by an algorithms called S++ algorithm which was also the basis for the discussed S# algorithm which was capable of cheap talk.

more to technology. Because the amount of interaction between humans and technology has intensified rapidly over the last decades, it seems reasonable to predict that this regular exposure will make humans more used to machines and more eager to cooperate.[4]
Also, we might suspect that in order to prevent bias towards algorithms, there is a need for more in-depth knowledge of the general public of how algorithms and intelligent machines work. If humans are to interact and cooperate ever more with intelligent machines, they need to be aware of the benefits and shortcomings of cooperating with machines. This is especially valid for experts such as medical practitioners who showed a very weak tendency towards accepting algorithmic advice.

If imperfect cooperation mostly arises due to the fact that the human is reluctant to interact with a machine in the same way he or she would interact with a human agent, a possible solution to algorithm aversion could be to make the human falsely believe that he or she is interacting with a human agent. Because this involves lying or at least misguiding the human, there are obviously ethical concerns to take into account. This ethical dilemma was called the transparency-efficiency trade-off.
It raises the question of when and how the true nature of a machine needs to be revealed. For instance, in written online communication, it is very hard for the human to know whether his or her interaction partner is a machine or human. Even in voice communication, language production by machines has become so natural that humans can be misled into wrongly assuming to be talking to a human being while talking to an algorithm. Google presented its product Duplex which can serve as a personal assistant and make a restaurant reservation over the phone. These real-world conversations were remarkably humanlike and Google received heavy criticism as people were afraid of not being able to tell the difference (cf. newspaper arcticles such as Bergen and News (2018). While Google announced that the Duplex assistant will disclose its machine nature, it becomes clear that the question of transparency is becoming important. Should humans always have the right to know the nature of their interaction partner? How does the lack of transparency change human behaviour? We might suspect that people will use impolite behaviour more often when they assume a hotline voice to be a machine as they cannot tell whether it is human or not.
However, because of these ethical questions, we should be reluctant to try and overcome algorithm aversion with a lack of transparency. There seems to be a more efficient solution as even without lying or misguiding the human, we might nudge a human into cooperation by simply making the machine appear more human.
In fact, it was observed that when machines affectively modulate the voice in a way that is intuitive for humans, this significantly improves team performance, and appropriate affect

[4] This exposure effect seems to be very different than the discussed trust discounting in the context of the perfectionist scheme. In both cases, it is the interaction with a machine that leads to a change in the human attitude towards cooperation with machines. However, the two scenarios happen on very different timescales. Exposure is a long-term process whereas the discounting we saw was towards an algorithm is in specific situation. Therefore, the discounting of automation after seeing machines err is not an argument against our suggestion that exposure helps against algorithm aversion.

expressions by the robot help humans to accept robot autonomy (Scheutz, 2011).
We shall therefore now turn out attention towards the human anthropomorphization of machines to be able to assess if and how algorithm aversion might be overcome by anthropomorphization.

Chapter 4

Humanizing machines

"Don't call me a mindless philosopher"

The humanoid robot C-3PO in the movie Star Wars, 1977

In this chapter we investigate the human tendency towards humanizing machines and we refer to both anecdotal and empirical evidence of aspects of anthropomorphization. This is relevant as we would like to understand if and how anthropomorphization might help solve problems of algorithm aversion in human-machine cooperation.

4.1 A working definition of anthropomorphization

In order to discuss the humanization of machines, we start by giving a brief discussion of the term anthropomorphization. Epley, Waytz, and Cacioppo (2007) have proposed a psychological framework to understand anthropomorphization which they define as "the tendency to imbue the real or imagined behavior of nonhuman agents with humanlike characteristics, motivations, intentions, or emotions" (Epley et al., 2007, p. 864). Here, we focus on the anthropomorphization of intelligent machines as nonhuman agents.

A similar idea is the concept of the intentional stance by Daniel Dennett (Dennett, 1971). Adopting the intentional stance means planning and predicting the behaviour of other agents with reference to their mental states. Regarding an intelligent machine, Dennett claims that "we find it convenient, explanatory, pragmatically necessary for prediction, to treat it as if it had beliefs and desires" (p. 91f.) Then, the intentional stance would allow more efficient interaction.[1]

In this work we will use the term anthropomorphization[2] and adopt a working definition similar to (Epley et al., 2007) which naturally includes the concept of the intentional stance.

[1] A study conducted with regard to whether the intentional stance might apply to machines suggests that it is indeed possible to induce adoption of the intentional stance toward artificial agents in some contexts. Many participants were somewhat biased towards the mechanistic stance, however (Marchesi, Ghiglino, Ciardo, Baykara, & Wykowska, 2019).

[2] The act of humanizing machines will be used synonymously to anthropomorphization.

It suffices for the scope of this work to understand anthropomorphization as the attribution of humanlike features to nonhuman entities, i.e. machines.

4.2 Unidirectional relationships and perceived agency

While it is without doubt interesting to investigate how the process of anthropomorphization works and what the key determinants of this process are, we will leave this for later (cf. Epley et al. (2007) as well as chapter 6). Here, we will look at different cases of anthropomorphization to argue that the human tendency to anthropomorphize is strong.

The first examples of humans humanizing machines appear in the case of chat bots. We already introduced the bot ELIZA where a virtual Rogerian psychotherapist asked simple questions. ELIZA was not yet a very sophisticated program, yet, her creator reported that he was startled by how emotionally involved the tester became with the machine and "how unequivocally they anthropomorphized it". He goes on: "Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room. Another time, I suggested I might rig the system so that I could examine all conversations anyone had had with it, say, overnight. I was promptly bombarded with accusations that what I proposed amounted to spying on people's most intimate thoughts." (Weizenbaum, 1976, p. 6)
Over the last decades bots learned to master much more complex conversations. In the case of XiaoIce, a modern Chinese chat bot, there was a reported maximum conversation session of over 29 hours (Shum, He, & Li, 2018). When analysing the content of the conversations, we still see how humans tend to tell bots their inner secrets.
It seems unlikely that a person would have long and intimate chat session with a bot if the person perceived it as a technical tool only. This is because the combination of a statistical learning algorithms with a phrase database do not seem to be a great reference for a good conversation partner. The intensity with which human conversation with bots are reported can only be explained with the strong human tendency to anthropomorphize the chat bots.[3]

Not only bots but also machines are suspect of being humanized. This personification of machines is often unidirectional and it is noteworthy how little machines have to contribute on their end to any relationship as (Scheutz, 2011) points out. He refers to some examples where this can be well observed.
A first example are robots whose task it is to defuse improvised explosive devices (IEDs). There is anecdotal evidence of soldiers forming intense unidirectional relationships with these machines. Garreau (2007) tells of a situation where a heavily crippled robot was struggling

[3] One extreme example of a non-human conversation partner could be a tree. Clearly, a human would not talk to a tree because of its vascular system for water distribution or the texture of the tree bark. Instead, people might seek friendliness, resilience and loyalty in a tree. It is the attribution of these humanlike personality features which makes the tree become a possible conversation partner.

to move forwards after losing some of its limbs while defusing bombs. The test was aborted by the commanding officer as he found the situation of the robot too inhumane to continue. In another situation, a soldier was quoted as saying about his unit's robot that it "was part of the team, one of us. He did feel like family".

With regard to robot companions, Kahn Jr, Friedman, and Hagman (2002) describe how owners of AIBO robo-dogs refer to their machine pet when posting online about their experiences. Frequently, the owners made affirmative references to the perceived agency of the machine (in 60 % of the participants) or to the machine's social standing (59 %). The category agency includes when owners referred to, e.g., the presence of feelings, intelligence or a unique personality in the robot. An example of an affirmative statement regarding the social standing is given by the authors as "I care about him as a pal, not as a cool piece of technology" (p. 632).

However, the most striking case for unidirectionality is probably the humanization of vacuum-cleaning robots. These robots have been sold for almost the last twenty years and are able to autonomously navigate an apartment and clean the floor. Sung, Guo, Grinter, and Christensen (2007) report the intense unidirectional bonds that some people form even with these very unemotional machines lacking humanoid features except for the occasional beep sound. People ascribed name, gender and personality to their Roomba robot, shared their experiences with others and were even willing to take up extra cleaning work in order to make the vacuum-cleaning robot function effectively.

Scheutz (2011) reports on experiments with regard to perceived machine agency and gives evidence that the degree of autonomy that a robot exhibits is an important factor in determining the extent to which it will be viewed as humanlike. Autonomous robots are machines that possess an autonomous skill such as free movement, object recognition, human-speech interaction or decision-making. Interestingly, even lack of effort in such robots makes robots appear to have agency (van der Woerdt & Haselager, 2019). The autonomy seems to be a critical factor in shaping human perception of the autonomous machines as having perceived agency. Scheutz (2011) concludes that this is experimental evidence that humans prefer autonomous robots over non-autonomous robots for collaboration.[4]

Generally, there seem to be two main drivers of anthropomorphization which are often both present at the same time. First, the nonhuman agent might have a humanlike appearance such as facial features which makes it easier for the human to socially connect to the machine. The second driver comes from the unpredictability of machine behaviour which makes the machine appear more autonomous (cf. van der Woerdt and Haselager, 2019).

[4] They also observed that robots can lead to social inhibition and facilitation effects merely by their presence. In one of their studies male participants showed a social inhibition effect during a math task. At the same time, the male participants viewed the robot as more humanlike than the female participants (Scheutz, 2011).

4.3 Complex and paradoxical aspects of anthropomorphization

4.3.1 Humanizing robots in the context of empathy and abuse

There is early evidence that humans react similarly to violence or torture towards humans or robots. In an fMRI study, Rosenthal-von der Pütten et al. (2013) used videos showing both positive and negative human interaction with the small dinosaur robot Pleo. Video clips with positive interaction included a person tickling, hugging or caressing the robot. For negative interaction, however, the dinosaur robot was, e.g., strangled, hit or captivated. These videos were shown to the participants while an fMRI scan of their brains was performed. The authors observed similar neural activation patterns for interaction with humans and robots indicating that human interactions with humans and robots are equally emotionally relevant for humans. They did measure a different neural activity in the right limbic lobe when comparing negative human-human-interaction with negative human-robot-interaction which indicated that human empathy is still larger with a human in a violent situation.

In a different study, Darling, Nandy, and Breazeal (2015) asked participants to observe a small robot toy Hexbug and then strike it with a mallet. They observed increased hesitancy when the robot was introduced using anthropomorphic framing that included name, personal features and a backstory.[5]

However, a clear picture of emphatic behaviour towards robots seems to be still lacking. For example, Cross et al. (2019) did not find evidence that a one-week socialization intervention with an engaging social robot led to a more humanlike empathic response to seeing this robot in pain.

Abusive behaviour towards robots is reported repeatedly and is the subject of empirical research. To protect robots from abuse, Brscić, Kidokoro, Suehiro, and Kanda (2015), for instance, developed a framework to prevent robots in a Japanese shopping mall to be kicked by children.

It is paradoxical to observe how the humanization of machines is accompanied by dehumanizing behaviour. We can only speculate about the causes but it seems that fears connected to emerging robot technology are turned against the machines. In most cases, the function of a robot is not fully understood by the humans interacting with it, and due to its passivity, the machine is a perfect victim. It was suggested that transparency and humanoid framing are effective in the promotion of human acceptance of robots and help to decrease abusive behaviour (cf. Bromwich, 2019).

[5] Darling et al., 2015 used two different backstorys. One backstory focussed on the personification of the robot: "This is Frank. Frank is really friendly but he gets distracted easily. He's lived at the Lab for a few months now. He likes to play and run around. [...]" The other backstory highlighted personal experience: "This object has been around the Lab for a few months now. If you had come by before, you would have seen it moving around on the floor. It gets around but doesn't go too far from the lab. [...]". The participants hesitated significantly longer to strike the Hexbug for both backstory conditions. There was no significant difference in hesitancy between the two backstory conditions.

4.3.2 The uncanny valley hypothesis

In the context of humanoid appearance, we include a brief glance at the perceived effect of anthropomorphic machines on humans.

It has been suspected for quite some time already that the relationship between human familiarity towards a robot and the human likeness of the machine is not straightforward. Instead, it is observed that if the robot only looks slightly like a human, but not quite so, then humans become quite sceptical of the robot. This was first described by the Tokyo robotics professor Masahiro Mori as the *uncanny valley hypothesis* where the uncanny valley describes the dip in the curve where close resemblance to a human is perceived negatively by many people Mori (1970).

This effect resonates well with personal experience. It seems much easier to attribute personal features to distorted characters with unnatural facial and body proportions in animated movies than to pseudo-realistic graphics found in some movies such as The Polar Express (2004), for instance. The uncanny valley hypothesis (UVH), however, is lacking consistent empirical support and cannot be generalised across different individuals, stimuli, situations, tasks, and time (Cheetham, 2017). Even though findings towards the UVH are inconsistent, it is referred to in many research papers. We should conclude that even subtle humanoid features can already have a strong impact on how humanoid a machine is perceived.

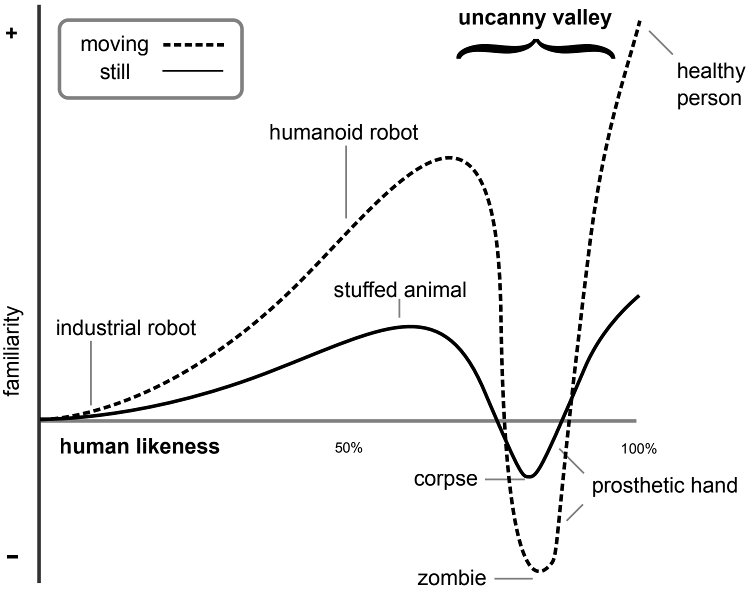

Figure 4.1 – Relationship between familarity and human likeness of a robot. The uncanny valley describes the dip in the curve for high human likeness. In the UVH, the effect is larger for moving robots than for still machines (adapted from Wikipedia contributor Smurrayinchester, 2007; based on Mori, 1970).

4.4 Conclusions on anthropomorphic framing

We investigated the most important aspects of anthropomorphization and confirm the human tendency to humanize nonhuman agents. When humans ascribe humanoid features to machines, they become more inclined to interact with machines. We saw, however, in the examples of robot abuse and the uncanny valley hypothesis that might lead to surprising and conflictual results in some cases. The fact that the process of anthropomorphizing machines is complex should make us careful in the application of humanoid framing. More research is needed into the circumstances and mechanism of humanoid framing. For that reason, we propose an empirical study in the last chapter of this work.

Still, the strong tendency for anthropomorphization is evidence that humanoid features in robots can help greatly promote the acceptance and the integration of intelligent machines in society.

Chapter 5

The anthropomorphization nudge for human-machine cooperation

We have seen that human features are helpful for human-machine cooperation and that there is a natural tendency for humans to humanize machines. A nudge is supposed to alter "people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." (Richard H. Thaler, 2008, p. 6) When a certain design of an intelligent machine is used intentionally to induce a certain human behaviour, then this is a form of nudging.

We are interested in this anthropomorphization nudge to improve the efficiency of human-machine cooperation. In this chapter we investigate the effects and consequences of anthropomorphic framing.

5.1 Observations on anthropomorphization of robots in various contexts

We will argue that anthropomorphization is useful and desired in some instances, however, in others it is not. In order to do so, we will investigate situations where the anthropomorphization nudge plays a beneficial role and also situations where anthropomorphization can pose a risk.

5.1.1 Humanoid features fostering acceptance of and interaction with machines

We saw in the previous chapter that humans prefer to cooperate with a machine partner that features autonomous and humanoid features. These features included humanlike speech or autonomous decision-making and lead the human to perceive the machine as having some degree of agency.

Our main reason why we evaluate anthropomorphization is to improve the efficiency of human-machine cooperation. However, benefits of anthropomorphizing machines go beyond overcoming algorithm aversion. Indeed, there are many cases where machines can be a catalyst for inter-human interaction. Humanoid robots have successfully been used with children who have autistic spectrum disorder, as a facilitator for doctor-child communication in hospitals, as a motivator for regular fitness exercise and as a valuable tool to engage children in learning (Darling, 2017).

A successful social robot is the baby seal Paro. There are a number of applications where Paro has been used to calm down distressed people and as a replacement for therapeutic animals. Paro is able to give humans a sense of empowerment and it has been used in nursing homes since 2004. There, it inspired more conversations and interaction among the residents (Kidd, Taggart, & Turkle, 2006). While the baby seal robot does have the appearance of seal instead of humanoid body, it still has humanoid features such as positive emotional reactions to being stroked. Paro is not perceived as a mechanical tool but as an animal with a personality.

These are examples of how social robots can facilitate communication between humans and be a welcome supplement to human interaction. We note that machines are able to and should always support the empowerment of humans (cf. Bracy, 2015; Darling, 2017, p. 177).

5.1.2 Anthropomorphization can promote undesirable behaviour

While anthropomorphization can support human-machine cooperation, there are also examples where undesirable behaviour is promoted. The example of the Anti-IDE[1] robot shows how anthropomorphization can prevent the machine from its intended use.

"Just as a human team would ’leave no man behind,’ for instance, the same sometimes goes for their robot buddies. When one robot was knocked out of action in Iraq, an EOD[2] soldier ran fifty meters, all the while being shot at by an enemy machine gun, to ’rescue it.’ " (Singer, 2009, p. 339)

Anti-IDE robots are being designed with the goal to protect soldiers and make their work safer. This is because human life is irreplaceable and worth protecting while the memory chip of an unconscious machine can be cloned easily to a new machine. The anthropomorphization of such a robot makes it harder for the soldiers who work with it to see it merely as a replaceable machine. This led to the situation of a soldier risking his or her own life to rescue a machine.

However, undesirable behaviour with intelligent machines does not only arise unintentionally as in the case of the Anti-IDE robots. When the aim of the creators of an intelligent machine was to exploit the human, then increased cooperation due to anthropomorphization can be harmful. An example for this are humanoid slot machines. In a study Riva, Sacchi, and Brambilla (2015) showed that more humanoid gambling machines were anthropomorphized more by the gamblers. The participants were more inclined to play with these anthropomorphized machines which would cause them to lose more money.

Intelligent machines where anthropomorphization is central are the emerging sex robots. Due to the intimacy of sexual relations, these machines are especially noteworthy also when considering the risks involved. The psychological depencence of a humans being on an intelligent

[1] IDE: improvised explosive device

[2] EOD: Explosive Ordnance Disposal

sex robot might make the human partner very exploitable. Creators of sex robots could use this dependency in unethical ways, e.g. for political campaigns or financial exploitation.[3]
An increased exploitation of humans by anthropomorphic machines seems indeed possible as empirical evidence shows how humans are reliably more truthful with robots (Scheutz, 2011). This aligns well with the anecdotal evidence for chatbots. For example, we saw how people would tell Weizenbaum's ELIZA bot intimate secrets very eagerly.

We have observed that the risks of anthropomorphization fall mainly into two categories. First, anthropomorphization may prevent the intelligent machine from fulfilling its design purpose. Secondly, if a machine is designed to exploit humans, this will be facilitated by the use of humanoid features.

5.2 Ethical considerations of anthropomorphization as a nudge

We have seen that there are many scenarios where the anthropomorphic nudge plays an important role in human-machine cooperation. However, in several cases humanoid appearance promotes undesirable human behaviour. This is highly troublesome as intelligent machines are ubiquitous and opaque and can have a profound impact on our lives.
Especially, we believe that the arguments provided so far are a clear case against any natural tendency for ever more humanoid robots. When business interests might promote anthropomorphic framing in more and more settings, the potential risks of humanoid machines could be ignored. It might even be necessary to impose legal regulations of humanoid features in robots.
On the other hand, we witnessed that anthropomorphization can improve the acceptance of intelligent machines and promote human-machine cooperation. These benefits of humanoid features can significantly improve the performance of hybrid human-machine teams which will save human lives by better medical care, fewer road accidents, etc. Waving these possible benefits entirely by abstaining from the use of humanoid features would be clearly too restrictive.

We therefore argue for the encouragement of anthropomorphization on a use-case basis, similary to Darling (2017). The design of an intelligent machine must always also take into account the possible consequences of anthropomorphic design. Humanoid features should only be used when they are useful for the function of the machine. They should not be employed in cases where their effect is at best unclear or runs counter to the purpose of the machine. Also, sensitive applications such as gambling should not make use of the anthropomorphization nudge due to the discussed exploitation issue. We argue for an appropriate use of humanoid features in intelligent machines. Powerful technology needs to be designed

[3] Besides, sex robots likely will not care when they are being mistreated and we could speculate whether this could promote unconsented behaviour and misogynistic beliefs in society (Gutiu, 2016). This would be an example of the dehumanization of robots which can accompany their anthropomorphization (cf. section 4.3.1)

responsibly and this includes the responsible use of intended anthropomorphization.

However, there is the question of whether there could be spill-over effects from the anthropomorphization of some machines to those machines that should be perceived as tools only.
In our use-case-based scenario, we want to clearly distinguish two types of situations. Because humanoid features increase cooperation with machines, we want to use the anthropomorphization nudge in suitable applications only (context A), but we do not want humanoid features in applications (context B) where, for example, they might make humans exploitable. Hence, we want to promote anthropomorphization in context A and absence of anthropomorphization in context B.
Accordingly, what we do not want is anthropomorphization in context B and absence of anthropomorphization in context A. The first case is harmful in the sense discussed above while the latter case is the algorithm aversion inefficiency. We need to address a possible caveat that spill-over effects from anthropomorphization in context A might also lead to anthropomorphization in context B. These spill-over effects are imaginable if humans perceive non-humanoid robots as more anthropomorphic after being exposed to humanoid robots in their daily routine. Then, this spill-over effect could undermine the described use-case scenario.

5.3 Can humans distinguish between anthropomorphized machines and machine tools?

We would like to investigate how well individuals are in fact able to distinguish between machines that they humanize and those that they do not. We are concerned about a possible spill-over effect which could lead to an increased cooperation with non-humanoid machines after being subject to the anthropomorphization nudge with humanoid machines on other occasions. As we are unaware of any empirical studies that have tried to answer this question, we will look for analogies both in society and with regard to animals. In these examples, humans exhibit strikingly different behaviour towards individual human beings or individual animals, respectively, depending on the framing that was given to these individuals.
In history there are a number of distressing cases where humans were regarded differently with respect to their intrinsic value of being human. Severe marginalization, discrimination and persecution have occurred in many cases by constructing an in-group-out-group setting, constructed usually on the basis of national, religious or racial differences.
In the Roman Empire, slaves were not endowed with the same rights that the Roman people had, instead they were considered mere property. Similarly, racial segregation as in the well-known cases of the U.S. or South Africa show how a big part of the population was able to accept and internalize the artificial othering of people of colour.
Even though single individuals understood the cruelty of this power relationship, many others were able to internalize the arbitrary differences constructed between humans and humans stripped of their rights.

Another example for irrational compartmentalisation is found in the speciesism humans exhibit on a daily basis (cf. Norcross, 2004). The choices consumers make when buying meat products do not reflect rational behaviour about which species of animals they consume. Most people care deeply about domesticated animals such as dogs, cats or horses, but are very capable of ignoring the conditions which animals for meat production have to endure. It might have been reasonable from an evolutionary perspective to not eat a domesticated dog that is guarding you, but today a meaningful argument why people eat the animals they eat is usually lacking. The distinction in what animals are being used for food production is artificial as it does not follow a consistent rational, such as animal intelligence or health effects for the consumer.

The challenge of distinguishing humanoid intelligent machines from non-humanoid machines seems to be much smaller than the arbitrary and often cruel distinctions mentioned above. Therefore, we arrive at the conclusion that these examples seem to suggest the feasibility of an anthropomorphization nudge on a use-case basis.

Chapter 6

A study proposal on anthropomorphic nudging in the real world

In this chapter we will present the outline of an empirical study which is supposed to test our argument for anthropormorphization on a use-case basis. We have argued that human anthropormorphization of machines can promote human-machine cooperation. We saw, however, that a precondition for this is the human capability to distinguish the intelligent machines we humanize and those we do not.
We are unaware of any empirical studies that may have already investigated this question. We will therefore suggest to test whether a human participant will cooperate differently with a non-humanoid machine partner after he or she has cooperated with a humanoid machine partner before.

There are hypotheses which we have argued for above and that we would like to test in the empirical study. First of all, we have been arguing that humans have a greater tendency to cooperate with machines when they humanize the machines. This was an important point which we used when we argued for anthropomorphization for human-machine cooperation. We therefore want to confirm this in the experiment:

H1: Human cooperation with a machine partner is stronger when the machine is more anthropomorphized.

In the previous chapter we also argued that human behaviour in other contexts (racial discrimination and speciesism) seems to suggest that in fact it should be possible for humans to artificially perceive intelligent machines in different categories. There will be the category of machines which are supposed to being anthropomorphized by humans, here called humanoid machines, and there will be the category of non-humanoid machines being treated as mechanical tools. Then, we assume that it does not have any significant influence on human cooperation with such non-humanoid machines whether the human agent has had any exposure to a humanoid machine beforehand:

H2: Anthropomorphizing certain machines has no effect on the human cooperation with a non-humanoid machine.

In the study we seek to find evidence for these two hypotheses.

The aim of the proposed study is first to find evidence for H2. We suggest to use extreme cases of a humanoid machine which combines many features known to promote anthropomorphization and another non-humanoid machine without any of these. It should be noted that this implies that we will not be able to pinpoint findings to single characteristics such as the name or shape of the machine. Through the contrasting framings of the two machines we want to show that humans treat the two machines in very differently ways and only humanize the humanoid machine partner.

The humanoid machine will have both direct humanoid features[1] as well as features that promote perceived agency (cf. section 4.2). Regarding direct humanoid features we propose indicated joints such as arms and especially the outline of facial features. Also, communication is via voice communication similar to human natural voice in voice assistants of modern phones. In the light of the discussion on perceived agency, the humanoid machine should be able to move around and have some degree of autonomy in the decision-making process. Moreover, the instructions of the study will enable us to provide the humanoid machine with the right framing.

Humanoid framing gives a backstory to the machine. In our case, the story of the humanoid machine might be the following. It is presented as Alex, who will do everything to help you, the participant, in every possible way. Alex is extremely smart and very knowledgeable. Besides, Alex knows some of the best jokes and will never let you down. No matter how hard the problem seems, you can always trust Alex' friendliness and great team-work. Alex has already been living in our lab for more than four months and everybody in the team loves him, therefore we are sure that the two of you will also be a great match.

In contrast, the non-humanoid machine lacks all of these features. It is a small metal box which communicates using an unmodulated computer voice that lacks the naturalness of human voice. This non-humanoid machine is called Z2100 in contrast to a human name. While it does not have a backstory that promotes anthropomorphic framing, it is also introduced as being equipped with a state-of-the-art artificial intelligence and proved to be very reliable and useful over the last four months. Besides, Z2100 features cooperation skills on the level of human cooperation and its computational power will be a great asset to you, the participant.

We suggest to use human-human cooperation as control for the study. The control group will serve two purposes. First, we might get additional evidence regarding H1. We expect humans to be more inclined to cooperate with humanoid machines than non-humanoid machines, but less than with other humans. Using the human-human control scenario, we will be able

[1] The robot does not need to resemble a human very closely to promote anthropomorphization. We have discussed earlier that there is a natural tendency for humans to adapt and also because of the uncanny valley hypothesis we should be careful not to make the robot appear very similar to an actual human being.

to check whether this holds true. Additionally, we do not expect cooperation-interaction with another human to have an effect on the inclination to cooperate with a non-humanoid machine afterwards.

Therefore, we propose a study design where participants[2] are split randomly into three different groups. Group A will play a strategic game with the non-humanoid machine Z2100, Group B will play the same game with the humanoid machine Alex, and Group C as control will play the game with another human being. Afterwards, all participants will play a second instance of the game with the non-humanoid machine Z2100.

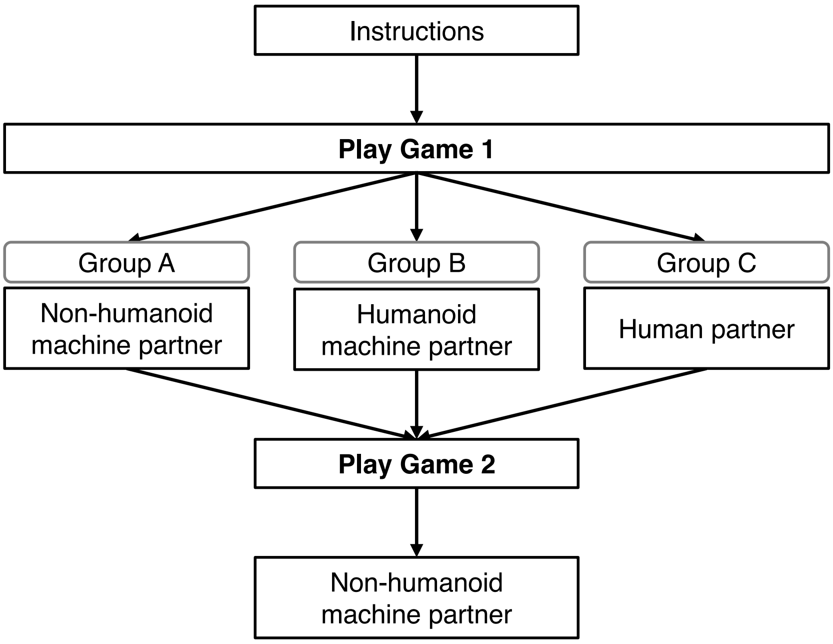

Figure 6.1 – Proposed study design for testing whether humans are able to adapt their behaviour depending on the amount of humanoid features present in their cooperation partner. Study participants are divided randomly into the three groups A, B, C. Each group plays against a different interaction partner in a first game instance. Then all participants play against a non-humanoid machine partner in the second game instance.

We will test the inclination for cooperation in the games. It is important to choose a game where both players benefit when they cooperate, but selfish or mean game-play should also be possible. A turn-taking, extensive game seems a reasonable choice. Therefore, we suggest playing the Block Game which Ishowo-Oloko et al. (2019) used in their experiment or a similar game.

[2] Using a power analysis, it is possible to estimate the optimal number of participants. This is because a too low numbers of participants increases the risk of not being able to come up with significant results.

A survey after the second round will provide retrospective insight into the question of how the participants felt about the situation. We propose to include a questionnaire on anthropomorphization such as the "Godspeed" questionnaire (Bartneck, Kulić, Croft, & Zoghbi, 2009) or the IDAQ (Waytz, Cacioppo, & Epley, 2010). The perceptions of the participants might provide further information regarding the reasons why and if humans distinguish between machines when anthropomorphizing them (H2).

In the study, we might want to control for different variables as they can influence the possible confirmation of the hypotheses. In section 4.2, we mentioned differences in social inhibition between male and female participants due to algorithms. Therefore, it is reasonable to introduce gender as a first control variable.
Furthermore, we might suspect that exposure to technology changes the way we interact with it. Younger people grew up with ubiquitous technology while older people have not. It seems speculative whether this would make it harder or easier for young people to distinguish between machines they humanize and machines they do not. However, as an effect on H2 seems likely, age is another reasonable control variable.
Additionally, it is unclear if the distinction (H2), which we want to confirm, requires additional mental capacity from the participants. If this was true, intelligence or working memory capacity of the participants can have an influence and might be considered as control variables.
There are possible extensions of the study. Instead of two games with only one round, we could alter the study design and play many rounds in both game instances. This might make the findings more robust and also provide insights into the development of cooperation over the rounds.
After confirming H2, we will be interested in the question as to which humanoid factors were crucial in creating the distinction between the humanoid and the non-humanoid machine. We might therefore repeat the experiment with many different versions of a humanoid machine. One version might only have a human name, another one only human shape or only a human backstory. It seems interesting which of the factors or which combination of them is most important.

We are confident that the presented study proposal will help to build the argument for the anthropomorphic nudge in human-machine cooperation. If humans are able to distinguish between humanoid and non-humanoid machines, anthropomorphization could be used to promote human-machine cooperation while minimizing negative side effects.

Chapter 7

Conclusion

Intelligent machines have a profound impact on individuals and society. We showed how human-machine cooperation plays an important role in the transformation that is taking place and should therefore be a key focus of machine behaviour. In this context, we discussed algorithm aversion and the human reluctance to cooperate with (intelligent) machines.
There is a transparency-efficiency trade-off regarding human-machine cooperation as misguiding humans into believing that their cooperation partner was human increases efficiency. We argued that the human tendency to anthropomorphize machines can be used to avoid the ethical difficulties that would arise from a lack of transparency. Anthropomorphizing behaviour can be fostered through humanoid and autonomous features in machines. Influencing human interaction with a machine through these humanoid features is the anthropomorphization nudge.

However, there is a downside connected to the anthropomorphization of machines. It might promote undesirable behaviour in humans and make them more exploitable. We thus argue that machines should bear humanoid features only if there is a legitimate reason for this. The question when humanoid features are used should be made on a use-case basis.
A precondition for this use-case-basis scenario is the human ability to distinguish between humanoid machines which are being anthropomorphized and non-humanoid ones which shall not. We hypothesize that humans have this ability and suggest a study design which is able to test this hypothesis.
With a confirmed hypothesis, we advocate the anthropomorphization of certain intelligent machines to increase the efficiency of human-machine cooperation. We are that this framework may enable humans and intelligent machines to cooperate in a way that is beneficial to humans and to society.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283). Savannah, GA: USENIX Association. Retrieved from https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi. (accessed 07 Jan, 2020)

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*(1), 71–81.

Bergen, M. & News, B. (2018). 'Silicon valley is ethically lost': Google grapples with reaction to its new 'horrifying' and uncanny ai tech. Retrieved from https://business.financialpost.com/technology/personal-tech/silicon-valley-is-ethically-lost-google-grapples-with-reaction-to-its-new-horrifying-and-uncanny-ai-tech. (accessed 07 Jan, 2020)

Boden, M. A. (2018). *Artificial intelligence: A very short introduction*. Oxford University Press.

Bostrom, N. (2016). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bracy, J. (2015). The future of privacy: My journey down the rabbit hole at SXSW. *Privacy Perspectives*.

Bromwich, J. E. (2019). Why do we hurt robots? The New York Times. Retrieved from https://www.nytimes.com/2019/01/19/style/why-do-people-hurt-robots.html. (accessed 07 Jan, 2020)

Brown, N. & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, *365*(6456), 885–890.

Brscić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from children's abuse of social robots. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 59–66). ACM.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., . . . Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Apollo - University of Cambridge Repository. Retrieved from https://www.repository.cam.ac.uk/handle/1810/275332. (accessed 07 Jan, 2020)

Cheetham, M. (2017). The uncanny valley hypothesis and beyond. *Frontiers in Psychology*, *8*, 1738.

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., ... Rahwan, I., et al. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233.

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180034.

Danks, D. & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Darling, K. (2017). ’Who’s Johnny?’ Anthropomorphic framing in human-robot interaction, integration, and policy. *Robot Ethics 2.0*, 173–188.

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. In *24th IEEE International symposium on robot and human interactive communication (RO-MAN)* (pp. 770–775). IEEE.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Garreau, J. (2007). Bots on the ground. *Washington Post*, *6*.

Google. (n.d.). Basic classification: Classify images of clothing. Retrieved from https://www.tensorflow.org/tutorials/keras/classification. (accessed 07 Jan, 2020)

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Gutiu, S. M. (2016). The roboticization of consent. In *Robot law*. Edward Elgar Publishing.

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521.

Kahn Jr, P. H., Friedman, B., & Hagman, J. (2002). I care about him as a pal: Conceptions of robotic pets in online aibo discussion forums. In *CHI’02 extended abstracts on human factors in computing systems* (pp. 632–633). ACM.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics*, *01*(03), 465–480.

Kidd, C. D., Taggart, W., & Turkle, S. (2006). A sociable robot to encourage social interaction among the elderly. In *Proceedings IEEE International conference on robotics and automation. ICRA 2006.* (pp. 3972–3976). IEEE.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems 30* (pp. 656–666). Curran Associates, Inc.

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. Viking.

Leavy, S. (2018). Gender bias in artificial intelligence. In *Proceedings of the 1st international workshop on gender equality in software engineering - GE '18*. ACM Press.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., & Schellnhuber, H. J. (2019). Climate tipping points — too risky to bet against. *Nature*, *575*(7784), 592–595.

Levesque, H. J. (2018). *Common sense, the turing test, and the quest for real ai (The MIT Press)*. MIT Press.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Madhavan, P. & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Marchesi, S., Ghiglino, D., Ciardo, F., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance towards humanoid robots? *Frontiers in psychology*, *10*, 450.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Metz, C. (2017). In two moves, alphago and lee sedol redefined the future. Conde Nast. Retrieved from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/. (accessed 07 Jan, 2020)

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*, 33–35. Retrieved from https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley. (accessed 07 Jan, 2020)

Müller, V. C. & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 555–572). Springer.

Norcross, A. (2004). Puppies, pigs, and people: Eating meat and marginal cases. *Philosophical perspectives*, *18*, 229–245.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. (accessed 07 Jan, 2020)

Perez, S. (2019). Report: Voice assistants in use to triple to 8 billion by 2023. TechCrunch. Retrieved from https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/. (accessed 07 Jan, 2020)

Peter Eckersley, Y. N. et al. (2017). EFF AI progress measurement project. Electronic Frontier Foundation. Retrieved from https://www.eff.org/de/ai/metrics. (accessed 07 Jan, 2020)

Prahl, A. & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., . . . Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Richard H. Thaler, P. C. R. S. (2008). *Nudge: Improving decisions about health, wealth, and happiness* (1st ed.). Yale University Press.

Riva, P., Sacchi, S., & Brambilla, M. (2015). Humanizing machines: Anthropomorphization of slot machines increases gambling. *Journal of Experimental Psychology: Applied*, *21*(4), 313.

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Hoffmann, L., Sobieraj, S., Maderwald, S., . . . Brand, M. (2013). Neural correlates of empathy towards robots. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* (pp. 215–216). IEEE Press.

Russell, S. & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Addison Wesley.

SAE International. (n.d.). SAE international standard J3016. Retrieved from https://cdn.oemoffhighway.com/files/base/acbm/ooh/document/2016/03/automated_driving.pdf

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv: 1708.08296 [cs.AI]

Savage, N. (2019). How AI and neuroscience drive each other forwards. *Nature*, *571*(7766), S15–S17.

Scheutz, M. (2011). 13 the inherent dangers of unidirectional emotional bonds between humans and social robots. *Robot ethics: The ethical and social implications of robotics*, 205.

Shum, H.-Y., He, X.-d., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26.

Singer, P. W. (2009). *Wired for war - Robotics revolution and conflict in the 21st century*. Penguin Press.

Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Pelican.

Sung, J.-Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). "my roomba is rambo": Intimate home appliances. In *International conference on ubiquitous computing* (pp. 145–162). Springer.

The Editors of Encyclopaedia Britannica. (n.d.). The mechanical turk: AI marvel or parlor trick? Encyclopædia Britannica, inc. Retrieved from https://www.britannica.com/story/the-mechanical-turk-ai-marvel-or-parlor-trick. (accessed 07 Jan, 2020)

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460.

van der Woerdt, S. & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, *54*, 93–100.

Wagner, J. (n.d.). GDPR and Explainable AI. Retrieved from https://www.zylotech.com/blog/gdpr-and-explainable-ai. (accessed 07 Jan, 2020)

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.

Weizenbaum, J. (1966). ELIZA–a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W.H. Freeman and Company.

Wikipedia contributor Smurrayinchester. (2007). Uncanny valley. Retrieved from https://commons.wikimedia.org/wiki/File:Mori_Uncanny_Valley.svg. (accessed 07 Jan, 2020)

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig verfasst zu haben und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel benutzt zu haben.

München, den 12. Januar 2019