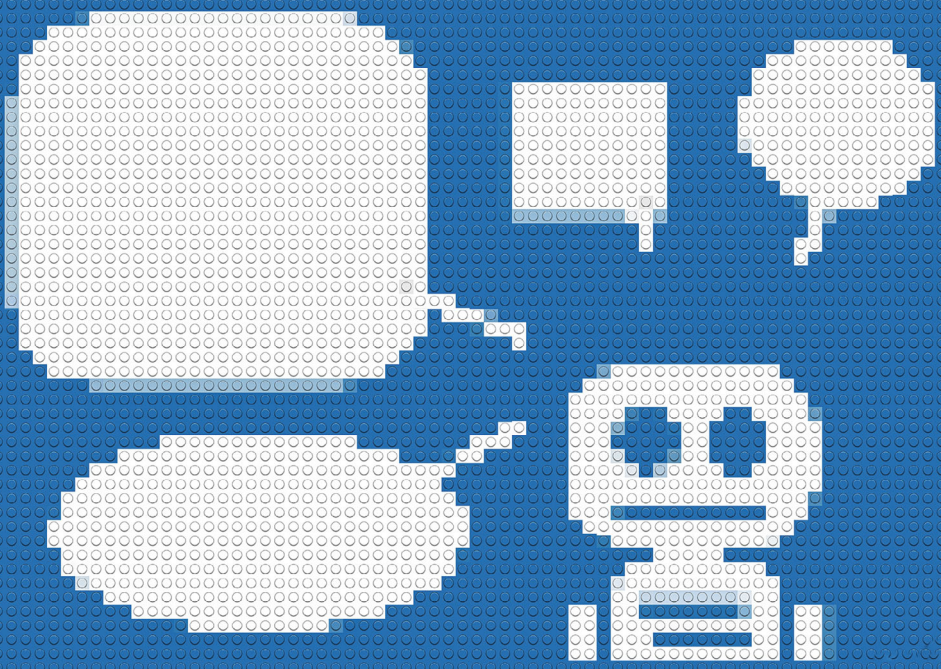
SOCIAL AND CULTURAL STUDIES OF ROBOTS AND AI



Chatbots and the Domestication of AI

A Relational Approach

Second Edition

Hendrik Kempt

palgrave
macmillan

Social and Cultural Studies of Robots and AI

This is a groundbreaking series that investigates the ways in which the "robot revolution" is shifting our understanding of what it means to be human. With robots filling a variety of roles in society—from soldiers to loving companions—we can see that the second machine age is already here. This raises questions about the future of labor, war, our environment, and even human-to-human relationships.

Hendrik Kempt

Chatbots and the Domestication of AI

A Relational Approach

Second Edition

palgrave
macmillan

Hendrik Kempt
Applied Ethics Group
RWTH Aachen University
Aachen, Germany

ISSN 2523-8523 ISSN 2523-8531 (electronic)
Social and Cultural Studies of Robots and AI
ISBN 978-3-031-75557-6 ISBN 978-3-031-75558-3 (eBook)
https://doi.org/10.1007/978-3-031-75558-3

Cover illustration: sakchai vongsasiripat/Gettyimages

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface to the Second Edition

When the first edition of the book was published in 2020, I had good reason to project confidence onto future developments of chatbot-technology. GPT-2 was just receiving its first philosophical attention, with philosophers exhibiting an uncommon urgency to react to this unprecedented jump in technological sophistication. This urgency, seemingly, stemmed from the fact that this technology behaves eerily close to our own ways, mostly by producing texts in all sorts of human ways of creativity and skill.

Since then, the term "Large Language Models" (LLMs) was added to many people's vocabulary, both academics and laypeople alike, as it appeared that the breakthrough has arrived to make AI work in relational capacities of personal and public use.

Then, in November 2022, after some months of rumors about the next generation of language models, GPT-3 and its associated user-interface (i.e., a chatbot) chatGPT was released by OpenAI, after entering an exclusive deal with Microsoft. This announcement supercharged the development of AI-tools (and with that an entirely new industry), putting pressure on other large data-driven corporations like Google and Meta to publish their own LLMs. This was attested by problematic releases of Google's Bard and Meta's Galactica, the publication of Meta's LLM as open source, and the overly ambitious implementation of these models in their respective suite of products. In March 2023, openAI released GPT-4, an even more powerful LLM seemingly capable of abstract reasoning and other "emergent properties."

This initial exponential increase in capacity triggered a number of experts demanding a pause in the development, fueling the public's and policy makers' worries about further improvements that could be beyond what we wished for. The pause did not occur, but neither has the exponential increase in capacity materialized. All the while, most use-cases of LLMs do not seem to be reliable and safe enough to warrant widespread adoption in legally sensitive contexts.

Next to the long-expected AI-hype triggered by the release of chatGPT, some of the previously speculative philosophical points seem to have been confirmed, and some new ones have arisen. The concept of "hallucinations" has been an example of the lack of reliability, the prowess contained in these LLMs has been shown to be easily harnessed for impermissible purposes through jailbreaks, and the resource consumption of LLMs and their respective data centers has upended the tech-sectors climate goals.

My work in this matter has been active and evolving. Thus, a second edition of this book was called for, as my basic idea of how we ought to relate to this new technology, and how we should deal with it moving forward, still stands. In the meantime, I have worked on ethical issues surrounding the use of AI, ranging from explainability concerns (Kempt 2024), the sustainability of AI (Heilinger, Kempt & Nagel 2024), the ability to be friends with machines (Kempt 2022), and what these chatbots ultimately ought to be allowed to say (Kempt, Lavie & Nagel 2024). All these ideas have featured in how I view the current landscape of LLMs, other AI technologies, and their relational capacities.

The changes made in this book are thus twofold. On the one hand, updates and changes to claims regarding the technology were made wherever necessary to accurately reflect and incorporate the developments over the last four years. However, it is a fool's errand to correct one's own predictions about where the technology will stand in a few years. Thus, on the other hand, I have added two chapters, philosophically approaching the ethics and the relational quality of LLMs. This, I hope, presents a sensible addition to the book without re-writing its core message of the possibility to recognize AI as something relatable if we make it so.

Aachen, Germany Hendrik Kempt

Acknowledgements

I am still grateful for the discussions I have had with others about relational technologies and associated issues. First and foremost, my co-authors of some of the papers mentioned above: Saskia Nagel, Jan-Christoph Heilinger, and Alon Lavie. Their wisdom and engagement with the development of generative AI, and the ethical and relational challenges associated with these technologies, have made it possible for me to work on this topic with enthusiasm. Their understanding and curiosity to work with my perspective has also contributed greatly to my continued excitement about these topics.

This book, however, would also not have been possible without the countless other discussions I have had with my colleagues in Aachen and abroad. Niël Conradie, Camilla Colombo, W. Jared Parmer, Peter Königs, Frieder Bögner, Giacomo Figá Talamanca, Nils Freyer, and John Mendonça have all been helpful in a variety of ways for which I am most grateful. I also want to thank Niklas Michels and Henry Seidel for the administrative support.

The conversations I had working on the project VEREINT (funded by the Federal Ministry for Research and Education BMBF, funding number 16SV9111) helped me re-organize my thoughts about how connected and relationally psychology features in our interactions with and through technology.

I also owe my thanks to Rachael Ballard and Naveen Dass from Palgrave Macmillan and Springer, whose patience and understanding of my needs and ways of working are very much appreciated.

Lastly, I still am indebted to my family, my partner, and my friends. Thank you for your support, your open ears, open minds, and open hearts.

Contents

CHAPTER 1

Introduction

1.1 Introduction

When Sony in 1999 introduced the toy pet Aibo, it was a small sensation. Aibo could replace actual pets on a level of sophistication never seen before, with the ability to simulate needs, interact with humans autonomously, including cuddles. The latest version can even differentiate whether the owner is preparing its "food" and "squirms in anticipation" (Sony 2020) to complement the already convincing simulation.

Initially, people were concerned that children were not able to differentiate between these plastic-based toys and actual animals, possibly losing respect and sensitivity toward living and breathing animals. However, those fears did not materialize, as most children instead developed the same affection and respect for their plastic creatures as they had for animals and human beings.

This story is merely an innocuous anecdote about the introduction of a somewhat autonomous toy. However, it nicely demonstrates two points that motivate this book: first, many new technologies change our way of life. Second, the consequences of adopting new technology may take a very different turn than anticipated.

For some people, new technologies represent hopeful progress toward a better future. For others, the very same technologies are viewed as threatening to the way of life they are accustomed to. No matter whether

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_1

one views technological changes as net positives or negatives, the common ground is that technology facilitates change.

Not any kind of change. Technological progress most often results in *social* change. It changes the way we relate to our environment and to each other. New communication devices allow for constant interactions with people thousands of miles away, while augmented reality will add another layer of interaction and information to our immediate surroundings. Medical progress allows for curing diseases that just decades ago were death sentences, while the latest autonomous battle drone can strike without even being noticed by its target. Some social movements like the Arab Spring would not have been possible without ubiquitous social media, while equally allowing oppressive regimes to implement measures that are even more oppressive, like the Chinese social scores. Technological progress may result in social change, but it does not guarantee social progress.

New technology always comes with risks associated with its use, and in open societies, discourses about the acceptability of those risks determine the overall acceptance of such technology. Artificial intelligence has, since its appearance in public discourse through several breakthroughs, been the subject of considerable social, political, philosophical, and economic reflections. A venerable AI-hype has emerged surrounding Large Language Models (LLMs), the key technology of this decade. From the Fall of 2020 to Summer of 2024, several remarkable improvements have been witnessed with this technology. This speed, often likened to a technological revolution, has also opened up many philosophical questions that are just slowly being asked and subsequently answered. Some of those questions we will ask here, and hopefully provide some answers, too.

1.1.1 Smart Fridges and Other Reifications

For a philosophical inquiry into issues of AI, one main obstacle appears right at the beginning: how can and should we understand the otherwise opaque concept of artificial intelligence? Due to the vast range of methods and applications that all are claimed to incorporate and exhibit some form of intelligent behavior, an all-encompassing definition will lose any practical purpose to limit any inquiry.

Take as an example: a "smart fridge." Its intelligence-claim is based on the ability to scan items in one's fridge and preorder those that, according

to typical use based on someone's consumption profile, will be used up soon, or warn about expired articles in the fridge. Calling the fridge smart, then, is a reification of AI, as it is not the fridge as a whole, but the added software and its connection to the cloud that is providing the smart stuff.

We are used to identifying intelligent beings as embodied entities occurring in nature, and this phenomenological basis is often the cause for misattributions and confusion about the source of (artificial) intelligence. This observation suggests that we require a fundamentally different approach to artificial intelligence than to natural intelligence: AI may come disembodied or might be re-embodied, duplicated, changed, and adjusted to the tasks at hand. It never is just one intelligent artifact, but an algorithm capable of operating on other soft- and hardware. To some degree, this argument also applies to approaches in philosophy of AI that concentrate on robots as embodied forms of artificially intelligent agents. Researchers have long argued that embodiment is a prerequisite for many cognitive capacities (Duffy and Joue 2000). However, this does not mean that robots are to be considered the intelligent entity, but that they operate with an intelligent algorithm. We should recommend, then, that philosophers carefully define the object and scope of their inquiries to avoid reification. This requirement is helped by the recent association of "AI" with "chatbots": the largest progress made in terms of intelligence has been to fine-tune language models to be as versatile and context-sensitive as possible. Our current ideas of intelligence, seemingly, reside within language.

In this book, we will consider chatbots and their social impact. Chatbots are computer programs, usually based on language models, capable of analyzing and reproducing natural language, that is the language human beings use to communicate with each other. Most of those chatbots are not embodied an an entity, even though their application is certainly not limited to chatrooms or being personal assistants in mobile phones and at home, where they currently are used most often, and with increasing numbers.

The reason for seeking out chatbots out of all current uses of certain kinds of AI is the capacity to speak with humans in their own language. This simple fact differentiates this technology from every other AI so far, and from every other technology as a stable, interactive communication based on people's own language has never been achieved before.

Engineers working in the area of natural-language processing (NLP), tasked with improving those skills, have no other way to go then to

imitate human language use further, which ultimately results in deceptive copies of speaking robots that not only use human language but imitate human speakers (take, e.g., Google's Duplex from 2018 as an earlier sign of purely imitating skills (Leviathan & Matias 2018). The better engineers follow this task, the more dubious their product becomes.

The goal of chatbot-engineers, possibly erroneously thinking that this is the road to artificial general intelligence (AGI), was for a long time to pass the Turing Test. This allowed the gamification of the very idea of the Turing Test: Imitating human speech, rather than human intelligence, lead to human-sounding machines that did not exhibit the levels of intelligence associated with human speech. The now infamous "stochastical parrot"-diagnoses by Bender et al. (2021) framed the understanding of what is going on in these chatbots rather well.

Reflecting upon those processes from a philosophical perspective, then, is rightfully alarming. The very many ways of using this technology for questionable purposes only increases with every step of improving those chatbots in becoming more "human." However, it also seems insufficient to merely assess this technology according to a human-nonhuman binary, since such binaries remain uninformative.

Recall how many children reacted to their plastic pets: they exhibited the same care and empathy they would have given to an actual pet; never mind how misplaced this empathy was. It seems we ought to explore a bigger picture when it comes to relating to robots and chatbots who just learned to speak like us. Chatbots in the future may acknowledge and remind us of their non-humanness without losing their appeal and relatability, which mostly stems from their use of natural language.

The core diagnosis of this book is that we do not have a bigger picture available yet, and constructing one is the task of philosophers of technology in the twenty-first century. Several philosophers have taken up that quest, and have been a great deal of inspiration for this book as well: David Gunkel, Marc Coeckelbergh, Kate Darling, Josh Gellers, and others who work to provide normative and phenomenological guidance for philosophically recognizing machines as something else than along the binary of "human" or "tool."

This coming century will undoubtedly bring new ways of human beings relating to their technological surroundings. One element of this bigger picture is the idea that our social categories with which we describe elements of the social fabric are woefully lacking differentiation. This lack of differentiation is the reason why some people rejected chatbots as any

possibly relatable technology, similar to people who rejected the idea of relating to toy pets.

Those limited social categories are both driving the engineering goals and the fears resulting from this successful engineering. Thereby, by arguing for including some human–machine relationships as social relationships, we can hope that the goal of imitating human speakers and full-fledged anthropomorphization loses its appeal. Open interactions with chatbots, finding a fitting place for them in our everyday lives, and thereby accepting the possibility of social human–machine relationships, is the proposal for a bigger picture.

1.1.2 What's to Come?

For this program to work, some preliminary clarifications ought to be made here. First, some methodological points are in order. These are presumably not necessary for the philosophically educated reader, but possibly quite useful for readers from other disciplines and backgrounds. The difference between an analysis and a reconstruction, for example, will carry some of the weight for this project, and it would be helpful for readers to follow this point. Second, we require a better idea of how chatbots and LLMs work, what the main objective in creating them currently is, and why we are justified in assuming that their progress in sophistication and prowess, even when not achieved with the current paradigm, is merely a question of time, rather than possibility. This chapter, in turn, may be of more interest to those of a less technical background. Being aware that the current standard of programming autonomous systems is machine learning with vast amounts of data will recalibrate the expectations of the future of chatbots and thereby allow for an empirical base of this inquiry. However, as with any philosophy of technology, some speculation and extrapolation are required. Third, a relational approach is developed to lay the groundwork of understanding social human–human relationships. This relational approach includes human-pet relationships as a precedent for incorporating non-human agents into the social fabric by giving them a unique social category. It also includes purely online-based human–human relationships as evidence that physical proximity is no longer a requirement for meaningful relationships. Fourth and fifth, the transfer of this relational approach to human–machine relationships is presented. This transfer should be considered the core chapters as they present the arguments to incorporate chatbots into the social fabric by

establishing a new social category, akin to a second domestication. The sixth and seventh chapters face the consequences of such a move, by acknowledging that this position requires a stance on the debate on robot rights, human-based design, the human–human consequences of human–machine relationships, and currently emerging considerations of the use of large language models for communicative purposes.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models be too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Duffy, B., and Joue, G. (2000). *Intelligent Robots: The Question of Embodiment*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.6703 (Last accessed on August 27, 2024)

Leviathan, Y., and Matias, Y. (2018): Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. *Google AI blog*. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (Last accessed on August 27, 2024).

Sony (2020). *Aibo Food*. https://us.aibo.com/feature/food.html (Last accessed on August 27, 2024).

CHAPTER 2

Methods

2.1 Method and Orientation

Some introductory remarks about the philosophy of technology are required to permit a philosophical analysis of the phenomenon of chatbots. This chapter is intended to serve this purpose, even though the richness and variety of philosophical methods in approach technology already doom the prospect of presenting an uncontroversial view. However, in acknowledging the futility of trying to write an adequate methods-chapter on the philosophy of technology, we can play the game a bit: In discussing the concepts we require for this overall project, and how we derive them, we can avoid long fought out debates.

In consequence, this chapter does not touch upon some of the more interesting theories and debates, like the approaches of postphenomenology or critical theory, even though we can be sure that their perspectives help approach technology philosophically. Instead, we approach the philosophy of technology and technology itself from the perspective of our linguistic conventions and how describing and distinguishing technological phenomena influences the way we think and assess those phenomena.

And since technology as a whole is almost impossible to define neatly, we are required to limit ourselves somewhat in scope. Luckily enough, the philosophy of artificial intelligence is an active and established field

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_2

within the philosophy of technology. In this field, we can first distinguish philosophical approaches to AI by two main areas: one discusses the prerequisites of AI by discussing philosophical concepts within the context of AI, and the other discusses the practical consequences of applied AI. These areas are not discretely distinct, as some theories in the former influence judgments in the latter.

The concepts of "intelligence" or "agency," the problem of artificial minds and mental states, and the question of when machines deserve the attribution of "agency" are all prerequisite philosophical discussions that may be informed by technological advancements. However, we may remain skeptical whether these questions can ever be satisfactorily answered. As Gunkel (2022) shows concisely, all of these—if used to derive an insight into our relationships with them—suffer from three shortcomings: we have issues agreeing which of these concepts is the most important one, how we would define it, and how we would detect it. These are all philosophically controversial decisions to make such a "property account" work. Ideally, then, the stronger argument and a more coherent organization of the invested conceptual inventory prevails.

The other focus lies on the consequences of AI. These consequences usually pose ethical questions of how humans want to create their society in the age of unprecedented computing power and autonomous agents. From the question of sophisticated robots disrupting labor markets and posing questions on the future of work, over mass surveillance courtesy of self-learning and data-gathering algorithms to controversies of relating to robots in emotionally significant ways (as lovers or as friends), the revolution of AI will affect every person one way or another (and probably already has).

Both parts are sometimes disregarded as rehashing older philosophical debates within the context of an emerging technology that is less revolutionary or problematic than presented in such debates (Nyholm and Smids 2016; Beard 2019). And while some applications of AI are certainly not revolutionary or deserving of a subsection of the philosophy of technology (since there is no "philosophy of airplanes" either), the potential to affect most people's lives is certainly there.

To appreciate this potentiality, interdisciplinary discussions between philosophers, engineers, sociologists, business leaders, lawmakers, and others ought to be encouraged and facilitated. For such debates to take off, a shared understanding of everyone's methods and terminology is welcome, even though philosophers do not have the best track

record of providing sensible insight into their terminological customs. By introducing a straight-forward philosophical perspective in the following chapter, we can hope to provide a ladder for those less familiar with philosophical methods.

2.1.1 Concepts and Conceptual Analysis

The main methodological purpose of this project is to reconstruct the meaning, scope, and content of concepts regarding the social implications of certain AI-driven technologies. There is an important distinction between an analysis and a reconstruction, and this distinction will carry this book's way of arguing. Conceptual analysis, as used in contemporary analytic philosophy, outlines the meaning, scope, and contents of concepts through differentiation and contextualization (Margolis and Laurence 2019). The primary purpose of an analysis is to sharpen the language used to describe the world to avoid certain philosophical problems that are based on an incoherent use of certain terminology. Some seem to hope that by way of such a conceptual analysis, many philosophical arguments turn out to be mere misunderstandings or mistakes on the questioner's side: clarity in language will, so goes the underlying assumption, lead to clarity in the world. However, the assumption about certain rationality within the use of concepts itself is a sign that such an analysis is inherently normative (Wedgewood 2007).

While conceptual analysis is usually thought of as non-normative, as it limits itself to analyzing concepts with reference to our intuitions and uses of said concept (as a term meaning "to dissect" or "to break into pieces"). Following this premise, conceptual analysis provides some definitions of concepts in accordance with our intuitions.

However, the purpose of philosophical analysis is usually to provide tested terminology that captures the use and content of concepts adequately. Thereby, analysis always provides a certain normativity about the *adequate* use of concepts. To acknowledge this normative dimension, we may speak of "reconstruction" rather than analysis. A reconstruction can count as an analytic effort because, in it, we attempt to find reliable central meanings through distinctions, comparisons, contextualization, and decontextualization. However, a reconstruction also allows for constructive suggestions on what a term's use should be. Not only is such a reconstruction bound to certain requirements of coherence, but it also may incorporate changes within the use of the term for more

practical purposes. In the tradition of certain constructivist approaches of philosophy, like methodical constructivism (Janich 1997), those uses and changes of use are taken from the lifeworld (Lebenswelt), i.e. the world in which our everyday life is happening. The reference to often pre-theoretical uses with certain concepts allows for a reconstruction of terminology that is close to "normal language philosophy" without making itself subject to those very uses.

In reconstructing concepts within the philosophy of technology, then, we may not only gather the emerging conventions of certain terms as used by engineers to pump intuitions (see Dennet (2014) for an elaboration on the concept of "intuition pumps"). It is an open philosophical question of how much intuitions should count in defining and forming terminology. On the one hand, intuitions provide a helpful initial idea about the scope of a term, and we can get good mileage out of relying on intuitions for guidance in different areas of conceptual analysis. On the other hand, certain uses of terminology create intuitions about the correct uses. Thereby, philosophical reconstruction should not shy away from making suggestions about specific uses of terminology.

2.1.2 Description and Evaluation of Technologies

To paint with a very broad brush, one could summarize the methodological approach of the analytical ethics of technology by first "describing" the technology and then by evaluating that technology.[1] The "description" unfolds by either naming and analyzing the features of current, existing technologies, or by extrapolating trends and assuming features soon to be brought to use in technology. Second, the "evaluation" is to put said technologies to the test within normative frameworks by assessing them and proposing certain restrictions on the development, implementation, and consumption of said technology.

Some issues lie between those two categories and are sometimes discussed in a confusing overlap of those categories, for example, the question of whether a robot could ever fulfill the requirements of personhood. This question can be interpreted from a merely descriptive

[1] This is not meant to disregard other approaches of philosophy of technology, like postphenomenological, postmodern or critical approaches to philosophy of technology. I am not only convinced that there is not one correct way of doing any kind of philosophy, but that a variety of approaches allows for the debate to remain productive.

perspective (in what Gunkel (2022) calls the property-account): either some technology will eventually tick all properties of a given concept of personhood, or it will not. If it is not capable of reaching personhood, it might be caused by implied properties of concepts of personhood that are not replicable and have some biologistic assumptions about the possibility of the emergence of personhood. Or the judgment rests on some strong assumptions about the ability of human-created technology to ever tick all presented boxes, which appears to be a rather strong thesis about the abilities of human creativity (and has comparatively few precedents in history).

However, the question about artificial beings reaching personhood could also be seen from a normative perspective: maybe someone does not want to share the conceptual space of personhood with anyone else other than the entities they are familiar with. Defenders of that case will play a kind of cat-and-mouse game with the opposite side, where the metalinguistic negotiation (Plunkett 2015) about what the concept of personhood should entail will always be moved to avoid the latest technological achievement checking all descriptive boxes. This game can go on until there are boxes that cannot be reasonably checked because they cannot be reasonably demonstrated to have been checked: for example, the existence of a soul or otherwise obscure constitutive personal interior. Or, even more obvious, the insistence that current machines cannot obtain personhood in a legally relevant sense (cf. Marshall 2023).

Thereby, when answering a question about ethics of technology one is presented with two fronts: On the one side, the metalinguistic negotiations which are often fueled by hidden normative agendas, and on the other side, an actual normative debate after everyone has agreed on the terminological inventory of the debate at hand.

For most debates, those two categories of descriptive and normative methods to approach technology are mixed. Metalinguistic negotiations and conceptual engineering efforts are both normative, as they represent discourses about the way we should use words and concepts, and to a degree, descriptive, as those debates lay the common ground of the very things we want to debate about. Consensus about the necessary features of personhood, to stay in this example, is a normative achievement that pre-structures any debate we might have about whether a certain type of artificial intelligence (or any type, for that matter) will be able to reach "personhood," and whether this ability is relevant to the question of how we ought to treat that technology in the meantime. These consensuses

avoid normative evasion-arguments, in which goalposts are moved along the advancements of technology to a point where those goalposts are outside the playable field.

These elaborations serve to show two things. One, it is important to keep this distinction clear to avoid misunderstandings. The main achievements of philosophy have been, arguably, established by important distinctions that allowed for expanding the understanding of certain philosophical issues. One could argue that some of the philosophy's strongest disagreements have ultimately been solved by introducing well-placed distinctions that allowed for a reassessment of the core disagreements. One could read Kant's introduction of categories of perception as such a philosophical distinction that effectively ended debate between rationalism and empiricism.

And second, allowing for negotiations about the proper use of a concept opens up a methodological space that will be exploited and built-upon in the following. The debate about personhood shows that its biggest challenge is to provide a consensus about the concept of "personhood." At the same time, the question of whether technology does (or will be able to) check off the subsequently spelled out features is one of the precise descriptions, which will also affect the debates about other terminologies and their uses.

However, the conceptual space is vast and incomplete. Comparative linguistics has shown just how many languages approach the world in vastly different ways, from the Navajo language that has a circular time concept to languages without subjects. And with a methodology in place that recognizes the necessity for debating the meaning and correct use of terms in debates, it ought to also recognize the necessity for sometimes constructing genuinely new terminology. With Wittgenstein's assumption of language boundaries constituting boundaries of one's world (Wittgenstein 1922, Proposition 5.6), it a simple deduction to propose that if we expend our language through distinctions and opening new categories, we are also expanding the boundaries of our world. Plenty has been researched in the empirical validity of the linguistic relativity on human world perception (empirically discussed under the name "Sapir-Whorf hypothesis" (Hoijer 1954)). Still, the philosophically more interesting point we are pursuing here is the impact of these expanded boundaries on normativity and the ability to describe certain phenomena in a different way. Without the distinction between intentional and unintentional body movements, we could not differentiate between a murder, manslaughter,

and a fatal accident. If we were told that turning a switch will kill someone in the room next to us (and will do nothing else), and we turn the switch and thereby killing the person, then we have committed a murder. If we have no idea about the situation and mistake the switch for a light switch, then we cannot reasonably be accused of murdering someone. Without the invention of "intention" as a feature of describing behavior as action, we would not be able to tell the difference between two very different situations.

Distinctions are not an end in itself since not all distinctions are productive in illuminating normative issues. The long-held distinctions between man and woman in legal codices, or between races, have been grave mistakes to make normative distinctions where no normative difference was to be marked. Making descriptive distinctions as a means to allow for more detail in describing a situation, often is exploited to attach normative distinctions as well.

In the tradition of analytic philosophy, it seems appropriate in most situations to offer more distinctions rather than fewer, as those distinctions do often help to discover normative differences of situations previously unknown. However, it is important to keep in mind that distinctions like those made here are a matter of social philosophy. They may, however, be exploited to make normative distinctions where there are no reasons for such distinctions. The simple fact that two things are different from each other does not carry any normative weight. This mistake is one crucial part of the naturalistic fallacy, in which the erroneous belief that descriptive distinctions are doing normative work is being held.

2.1.3 Distinctions and Discovery in Philosophy of Technology

Our general methodological point is for concepts to emerge and invite debate about proposals of how to understand those new concepts. The underlying assumption here is that, especially in the philosophy of technology, genuinely new ways of describing a technology will allow for a genuinely new way of describing the problems associated with that technology or even identify new problems altogether, both conceptual and normative.

The importance of those open conceptual approaches in the philosophy of technology lies within the intense speed with which technological progress occurs. Often those technologies enter an unprepared general public with insufficient awareness of its consequences. Thereby, the

general discourse of technology assessment depends on the often inappropriate characterizations of marketing or engineering departments. The way we describe technology is partly predetermining our judgment of it. Thereby, keeping one's terminological approach open to change to describe an equally open field of technology is paramount to describe and assess the technology at hand in the first place.

However, two things ought to be pointed out regarding the required openness for conceptual changes. First, the latency of philosophical progress, often criticized by futurists and technology-advocates (and apologists) to be a hindrance to progress (Hafner 1999). This latency is a useful counter in the context of market-logic dominated engineering goals and related hypes and unjustified bubbles (akin to the one we might be in at the moment). With incentives of overpromising and overadvertising technology that remains unintelligible to laypeople and unsafe for widespread use, it often remains unclear whether a new technology, in fact, poses genuinely new questions relevant to philosophy. Some distinctions offered by engineers will not hold as useful distinctions but are rather reflective of the science fiction they employ to protect their long-term engineering goals. Some others, such as strong anthropomorphism when describing robotic behavior, can count as a merely careless approach to language, or influenced from a perspective of selling certain technology and decreasing the resistance to new technology. Exposing distinctions without differences, like the way some engineers describe the behavior of their robots in colorful anthropomorphic terms,[2] requires that those fake distinctions have been made in the first place. Advocating for awareness in the descriptors used does not obligate philosophers to be language police of the sciences, but rather the referees of scientific discourse, of which language is a part.[3] If a discipline introduces distinctions and new concepts, philosophical work lies in reconstructing their uses, scope, and content, and testing them for plausibility, stringency, definitions, and rationality.

[2] An example taken from Sony's website for selling their toy pet Aibo: "Give aibo food and watch as it digs in happily. You can enjoy meals at the same time or tell aibo to wait and watch it squirm in anticipation." (Sony 2020).

[3] The idea of scientific discourse boils down to rules of scientific arguments. The body of works regarding the variety of arguments in scientific discourse is vast. For an extensive overview van Eemeren et al. (2014) is recommended.

Second, to keep inter-philosophical debates coherent, philosophy of technology needs to keep a tether to some basic philosophical concepts. Thereby, it appears reasonable to assess new technologies and their new approaches to a certain practical problem with the established philosophical concepts to see how far such approaches carry. Without certain principles in assessing and describing technology, philosophy of technology would not provide any productive insight but would merely generate philosophical justifications of a given moral trend.

Another example from AI helps to illustrate this point: the Trolley-cases, brought into the broad debate among philosophers by Philippa Foot (1967), aim to invoke some intuitions about the normative relevance of actions vs. inactions, as well as the already normative relevance of the amount of damage dealt in a situation. According to these Trolley-cases, the difference between action and inaction, as a first descriptive distinction, is also being considered normatively relevant due to some intuition that "doing something" is normatively more relevant than "doing nothing" (for an extensive discussion of the "doctrine of double effect, to which the Trolley-cases allude to, see McIntyre (2018)).

However, a loose Trolley running down tracks and all the associated issues are, as Nyholm and Smids (2016) point out, thought-experiments, i.e., arguments of hyper-specific features that are supposed to isolate certain intuitions about those hyper-specific features. Yet, with autonomous cars entering streets, at least the arguments and distinctions made in the debate around Trolley-cases are of burning significance (see Keeling 2019). And even though there still seem to be open questions on what exactly autonomous cars ought to recognize as protection-worthy and the Trolley-cases discussion will not yield immediately transferable rules (with some even doubting whether they can at all, see Himmelreich 2018), the progress made in that field (for those knowledgeable of it) is to no small degree traceable to the extended discussion of the Trolley-case, beginning in 1967.

Thereby, introducing technology as posing genuinely new questions that require a new set of distinctions and terminology has the burden of proof. This burden of proof is fulfilled if the limits of current categories of describing and evaluating technology are surpassed. In the following, I argue that some applications of AI technology, namely chatbots, will do exactly that, by providing sophisticated artificial natural-language users which are not sufficiently described and evaluated by relying on the technology we have produced so far.

2.1.4 *Reaches and Limits of Philosophy*

Mere philosophical arguments cannot induce societal changes in understanding and attitudes toward technology. It would be a mistake to assume that the recommendations made in philosophical discourses would amount to public opinion. Philosophical discourse is not equal to public discourse. A philosophical project thereby is mainly an ideally well-thought-through collection of normative suggestions of improving the discourse by providing clear renderings of arguments and questioning preconceived notions of certain concepts.

Thereby, rules of philosophical discourse are usually assumed to be somewhat different, and the conclusions drawn are often not immediately practical. One could describe philosophical discussions as rational pressure chambers to test the intricacies and extremes of certain positions. The results are, therefore, only an indication of where society could go, not where it exclusively should go.

Such a (meta-)philosophical position about the self-positioning of philosophy in public discourse of an open society has been prevalent for almost all of philosophy's history. It supports a pluralist, liberal approach to the prescriptions of how to live in a society, without shying away from making substantial recommendations based on rational arguments.[4]

2.1.5 *Moral Philosophy, Morality, Ethics*

Most public discourse is normative. In public discourse, questions of how we should live, both in a collective sense and regarding each person's behavior, are debated among a variety of participants with a variety of perspectives and "opinions." Many answers to those questions offered in public discourse are supported by moral arguments and moral codes, representing the moral convictions of different parts of society.

[4] This is not to say that philosophical discourse cannot or should not be challenged. In fact, many philosophers held highly problematic views that they defended using their own argumentative methods. For example, Immanuel Kant has being both arguing for universalibility in his ethical theory, yet also for the supremacy of white people (Boxill 2017). Without constant reflection on one's own position in the hierarchy of power and distribution of privileges, many normative statements are in danger of merely reflecting one's own privileges and thereby reaffirming current power structures.

Those moral arguments lead to moral conflicts, i.e., to conflicts about the moral thing to do in a situation. Some people expect moral philosophers to decide those questions. However, moral philosophy is a different term for ethics, not for morality, and moral philosophers are not in the business of deciding morality. The difference between ethics and morality lies in the level of abstraction: morality is about guiding people's actions by giving them specific rules, while ethics discusses those rules, their merits, and justifications. Moral philosophers, in this sense, are not content-experts, but method-experts.

Take, for example, the difference between a moral code, say, Chuck Norris's Code of Honor (Norris 2020), and the moral philosophy of Immanuel Kant. Norris provides specific rules on what to do, how to interact with others, and what to aspire with your actions. Kant's deliberations about the principles of morality have led him to different versions of the Categorical Imperative (CI), which represents a function on whether the maxims of certain actions are universalizable as rules for everyone (Kant 1785). The CI may apply to Norris' rules, but it does not mean that Norris' found the only moral rules available.

When philosophers work in the normative sphere, they usually propose and test arguments on how to provide principles and theories that can function as normative guiding rails for individuals contributing to public discourse, i.e., they do ethics. Thereby, the role of philosophers in moral matters is not to propose or evaluate the "correct" moral answers, but to propose principles which allow every single moral agent to come to their own moral answers and defend them against objections.

If moral discourse is understood as public discourse and ethical discourse as philosophical discourse, then the expertise of moral philosophers lies in the latter, not the former.[5]

[5] The concept of "moral experts," then, is a dubious one. A positive answer to the question whether philosophers have a privileged access to moral answers qua their ethical reflections is based on the assumption that there are moral truths to be known and those truths are found in philosophical reflection alone. Whether there are any moral truths to be known or rather merely universalizable maxims of action, is a controversial question that cannot be sufficiently answered here. However, in order to pragmatically undercut this controversy altogether, we may borrow René Descartes' idea of a "moral provisoire" (Descartes 1988, Chapter 3), which recommends the acceptance that moral thought remains provisionary and open to review until those principles of moral truth have convincingly been argued for.

However, this position also infers that philosophers are not moral experts or authorities through their philosophical thought but merely the referees of moral discourse.

If morality is fought out in public discourse, then it should count as a collective, social effort that is open to constant review, debate, and criticism. In such public discourse, philosophers are best suited to organize this effort in reflecting on the language and rules of the discourse. This position is usually referred to as discourse ethics (Habermas 1983, 1992), but does not imply that any moral conviction is valid. The rules of discourses require agents to accept certain conventions that make it impossible to hold certain moral convictions reasonably. For example, the moral position that murder is good is impossible to hold for anyone who would object to being murdered for no reason (we can assume that it applies to almost everyone). Discourse ethics also do not hold that philosophers cannot have their own moral convictions. They do not stand on the sidelines as referees, pretending to be outside the discourse altogether, but they are also taking part in the public discourse, as they should (Arendt 1958). However, their position does not carry more weight for the sole reason that they are philosophers (in the opposite, for example, to the medical evaluation of a doctor, or the opinion of a baker on someone's baked goods).

In consequence, this project does not make any moral assertions. A lack of moral assertions does not mean that there are no normative consequences from the theory laid out here, as it touches upon several debates within moral philosophy. Additionally, as elaborated before, the recommendations of using certain terminology in a certain way ought to be understood as a normative claim as well. However, as those recommendations are based on organizing a discourse, they are not intended to decide the discourse. What is judged as morally good or bad is, ideally, decided by a public discourse about the specifics of how to live, with rules guiding the discourse to guarantee an outcome every rational agent can accept.

2.1.6 AI Ethics, AI Safety, AI Governance

Lastly, the term "AI ethics" may receive a few needed clarifications through these characterizations of morality and ethics. Under the term "AI ethics" fall all efforts to formulate norms, rules, principles, and theories to provide the basis for regulating the technology of artificial

intelligence. As AI has many unique features that set it apart from other technologies, some combining principles are deemed necessary to reign in the risks from unregulated development.

Thereby, the efforts undertaken in AI ethics are not to tell companies to specifically leave out certain features in their products or not pursue a development idea. Instead, AI ethics offers methods and guidelines that have been agreed upon by the general public after a participatory discourse and possibly some vote in a deliberative body (see for example the AI ethics guidelines of the European Union, EU Commission 2019, having lead to the formulation of the AI Act of the European Union [2024]).

As those rules and principles are ethical guidelines, they only lay out the ground rules for the moral discourse on whether certain apps ought to be developed, or certain business practices are morally defensible. The overall moral judgments are up to the public discourse.

The issue some philosophers have taken with this entire approach is the fact that those methods and guidelines are often superficial and unenforceable (Vincent 2019, Yeung et al. 2019), reintroducing power imbalances due to the uniformity of those practicing AI ethics (Hagendorff 2020) or beset with toothless demands that lend themselves to ethics washing (Heilinger 2022) or even green-washing (Heilinger et al. 2024). At the same time, it still gives the impression that a company has done everything right according to a rigorous ethics code.

Lastly, we ought to distinguish AI ethics from AI safety. AI safety is concerned with making AI more safe, more reliable, potentially more explainable or transparent to achieve more safe or more reliable AI applications. AI safety, thus, while a normative undertaking, rarely addresses and reflects the assumptions of the technology in question. AI ethics, in turn, ideally critically assesses the conditions and consequences of the development, and widespread implementation of a technology. This division of labor can lead to a disconnect between AI ethicists, often working in academia or NGOs, and the industry, which is home to most AI safety engineers, changing the discourse away from substantial ethical concerns (such as data-gathering practices, responsibility gaps, or sustainability issues) toward practical questions of implementation, where the desirability of a certain AI application is taken for granted.

2.1.7 Conclusion

Explaining one's philosophical method and approach when beginning an investigation seems like an important yet sometimes under-appreciated step. The danger when doing so is to lose some readers right at the beginning, as they may take issue with the methods proposed here. It must almost be impossible, then, for those readers to continue. However, we can hope that not too many controversial statements were made at this point, as the next chapters will make even more.

Our primary concern in this chapter was to loosen the self-imposed limits of analysis by proposing to reconstruct terminology instead. A reconstruction takes into account the way we commonly use certain terminology and proposes how a definition free of (too many) inconsistencies could look like.

Part of these reconstructions consists in analyzing the adequacy of given distinctions and, wherever necessary, introducing new distinctions to capture differences in meaning and use. This method may lead to a different approach of explaining and furthering certain debates. Still, if the philosophical discourse is anything, then it is the place to propose unorthodox approaches and learn from them.

References

AI Act (2024). https://artificialintelligenceact.eu/de/ (Last accessed on August 27, 2024).

Arendt, Hannah (1958). *The Human Condition*. Chicago: University of Chicago Press.

Beard, Simon (2019). "The Problem with the Trolley Problem." https://qz.com/1716107/the-problem-with-the-trolley-problem/ (Last accessed on August 27, 2024).

Boxill, Bernard (2017). "Kantian Racism and Kantian Teleology." In Zack, Naomi (ed.), *The Oxford Handbook on Philosophy and Race*. Oxford: Oxford University Press.

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Last accessed on August 27, 2024).

Dennet, Daniel (2014). *Intuition Pumps and Other Tools for Thinking*. New York (NY): Norten Publishing.

Descartes, Rene (1988). *The Philosophical Writings Of Descartes*, 3 vols., translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press.

Foot, Philippa, 1967. "The Problem of Abortion and the Doctrine of Double Effect," Oxford Review, 5: 5–15; reprinted in Bonnie Steinbock and Alastair Norcross (eds.), *Killing and Letting Die*, 2nd ed. New York: Fordham University Press, 1994, pp. 266–279; reprinted in Woodward, (ed.), 143–155.

Gunkel, D. (2022). "The Relational Turn: Thinking Robots Otherwise." In Janina Loh (Ed.). *Emotional Bonds with Robots*. Transcript Verlag. pp. 55–76.

Habermas, Jürgen (1983), *Moral Consciousness and Communicative Action*. Reprint 2001. Cambridge (MA): MIT Press.

Habermas, Jürgen (1992), *Between Facts and Norms*. (Zwischen Faktizität und Geltung) Frankfurt: Suhrkamp.

Hafner, Katie (1999), Between Tech Fans and Naysayers, Scholarly Skeptics. *The New York Times*. https://www.nytimes.com/1999/04/01/technology/between-tech-fans-and-naysayers-scholarly-skeptics.html (Last accessed on August 27, 2024)

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heilinger, J-C. (2022). The Ethics of AI Ethics. A Constructive Critique. *Philos. Technol*. 35, 61. https://doi.org/10.1007/s13347-022-00557-9

Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice* 21, 669–684.

Hoijer, Harry, ed. (1954). *Language in Culture: Conference on the Interrelations of Language and Other Aspects of Culture*. Chicago: University of Chicago Press

Jan-Christoph, Heilinger Hendrik, Kempt Saskia, Nagel (2024). *Beware of sustainable AI! Uses and abuses of a worthy goal Abstract AI and Ethics* 4(2) 201–212 https://doi.org/10.1007/s43681-023-00259-8

Janich, Peter (1997). Methodical Constructivism. *Issues and Images in the Philosophy of Science*, 173–190.

Kant, Immanuel (1785). *Grundlegung zur Metaphysik der Sitten*. Riga: Hartnack.

Keeling, Geoff (2019). Why Trolley Problems Matter for the Ethics of Automated Vehicles. *Science and Engineering Ethics* 26, 293-307.

Margolis, Eric, and Laurence, Stephen (2019). *Concepts. Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/concepts/ (Last accessed on August 27, 2024).

Marshall, B. (2023). No Legal Personhood for AI. *Patterns* 4(11). https://doi.org/10.1016/j.patter.2023.100861.

McIntyre, Alison (2018). *The Doctrine of Double Effect. Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/double-effect/ (Last accessed on August 27, 2024).

Norris, Chuck (2020). *Code of Honor*. https://www.ufaf.org/founder_page.htm (Last accessed on August 27, 2024).

Nyholm, Sven, and Smids, Jilles (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice* 19, 1275-1289.

Plunkett, David (2015). Metalinguistic Negotiations and the Methodology of Philosophy. *Inquiry* 58(7-8), 828-874

Sony (2020). Aibo Food. https://us.aibo.com/feature/food.html (Last accessed on August 27, 2024).

Van Eemeren, Frans H., Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, Jean H. M. Wagemans, ed. (2014). *Handbook of Argumentation Theory*. New York (NY): Spring.

Vincent, James (2019). "The Problem with AI Ethics." *The Verve*, https://www.theverge.com/2019/4/3/18293410/ai-artificial-intelligence-ethics-boards-charters-problem-big-tech (Last accessed on August 27, 2024).

Wedgewood, Ralph (2007). *The Nature of Normativity*. Oxford: Oxford University Press

Wittgenstein, Ludwig (1922). *Tractatus Logico-philosophicus*. Routledge: London.

Yeung, Karen, Howes, Andrew, and Pograbna, Ganna. (2019). "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing." In Dubber, Markus, Frank Parsquale, and Sunit Das (ed), *The Oxford Handbook of Ethics in AI*. Oxford: Oxford University Press.

CHAPTER 3

The Social Dimension

3.1 The Concept of the Social

As a term first coined by Emilé Durkheim in the 19th century (1985), the term "social" has been interpreted and defined in many different ways, with consequences for the entire social theory build around the concept. Durkheim himself considered the "social" as object of the science of sociology the collection of "social facts." However, hopefully without alienating too many sociologists, we can characterize the social by reference to its everyday use as representing the collection of human interactions within a population.

Those interactions can be described as a network of relationships of human agents. This approach to describe the "social" is useful, as it allows to characterize most social events through the network of those social relationships. However, it ought to be clarified what relationships are and what makes them social.

Two conditions to define social relationships appear relevant here: they need to be mutually consequential, and they need to be embedded in a broader social network. The consequentialist condition for relationships to be social is that the actions of one party in this relationship are affecting the other(s). Often this condition is defined as reciprocal. (White 2015) However, it appears that reciprocity may not be helpful, as some relationships are not reciprocal in the stricter sense (like an infant and

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_3

their parents). They are, however, consequential for all sides involved. The embedding-condition is relevant to remain aware that social relationships are being formed, described, and experienced in a wider social network, with given rules, customs, and terminology that condition any relationship. (Granovetter 1985).

In this definition, a famous singer and their fans are in a basic social relationship, as their relationship is mutually consequential: A fan buys a ticket to the concert, tells others about the artist, or shares news articles about them. The singer, in turn, performs publicly for their fans, addresses them collectively on social media, or produces art with their fans in mind. They are also operating in an embedded context, as their relationship is shareable with others and recognized by others.

However, we cannot expect the singer to be aware of every single one of their fans for this relationship to count as one between "fan and artist." The actions of fans and artists affect each other (even though to a differing degree, with different awareness, different care, different emotional stakes, and different intentions).

In opposition to "pure" relational theories like the "agent-network theory" (Latour 2004, 2005), I do not claim that society cannot be analyzed through reference to non-relational social facts. Without norms and institutionalized rules, many of our relationships would not be explainable. Thereby, social relationships are only an explanandum insofar as our purposes here are to explain the dynamics of society as a relational concept. Whether a methodological individualism or a system-theory-oriented holism is the correct approach to explain how social relationships interact and scale is an open question, to which the answer is of no immediate consequence for this project. The point of view for this approach is not one of sociological theory, but philosophical: The fact that humans build relationships with each other, that those relationships are recognized or dismissed or otherwise judged by others, that those relationships differ in strength, reciprocity, depth, and relevance are all reasons to take them as a fundamental feature of how people are living together.

The explicit purpose for this project is not to reinvent sociology, but to find a conceptual structure coherent with many mainstream theories of sociology that allows for characterizing the social in a way that will open the social sphere to arguments of increasing its inclusivity.

In fact, understanding the concept of the "social" as primarily one of relationships between individuals, no matter how those relationships come about, allows for the conceptual space needed for the thesis pursued in

this book. As we attempt to evaluate if and under which conditions robots may enter our social sphere, we now can ask more precisely: what kind of social relationships can we enter with robots, and can we describe those relationships with the categories and terminology that we have used to describe other relationships so far? That is: are human–robot-relationships possible that are of a genuinely new kind, or are they bound to be in the image of human–human relationships, or maybe some other form?

3.2 Toward a Digital Society

One important detail for the upcoming reconstruction of the way we characterize human–human relationships and the differences, limits, and possibilities of human–machine relationships consists in the way the internet and the digitization have changed some underlying assumptions of social interactions. From public discourse and the relationship between policymakers and the general public on platforms like Twitter to intimate loving relationships between two people thousands of miles apart, social media has made new forms of social interactions possible. The speed of change of social relationships, in public and in private, has led some people to claim that technology has moved too fast. However, considering the inability to react to technological challenges socially, one may think that humanity is ill-prepared due to the lack of similarly significant social progress. The fear, then, is that we are woefully unprepared for the new possibilities of social interactions. The split between the "public" and the "private" is not always free of overlaps, especially on social media, where public and private conversations are as present as in a physical public square. However, this distinction helps to point out the relevance of the technological progress achieved in the last 30 years, especially when it comes to the challenges of democratic discourse as well as private uses of technology. The following chapter argues for the relevance of digitization in fundamentally shifting the constitution of public discourse and private lives. The argument does not claim to be a full elaboration on the society in the digital age, as this would not only surpass the scope of this project but is a substantial project on its own. Instead, the following aims to show that there is a reasonable way of talking about a "digital society" as a genuinely new subsection of human society in which people interact with others, build relationships, receive information, generate opinions, and spend a significant amount of time. This subsection is possibly detached, as new norms, expectations, and roles have developed that are distinct

from previous social developments. These changes have been happening ubiquitously over the last 20 years, and their consequences on society as a whole remain unclear.

3.2.1 Public Discourse

One of the most consequential changes in social interactions over the last 20 years has occurred in the realm of public discourse (Bouvier 2018). Those consequences are so severe that we have only started to understand what they are and how they affect the future of liberal democracy. On the one hand, the ease with which vital information spread has led to some substantial social changes, as the "Arab Spring" has shown. Without access to social media, these diverse, multicultural movements could not have spread with the speed and force as they have. On the other hand, liberal democracies have seen an increase in division and partisanship, quite possibly due to the formations of information and filter bubbles, echo chambers, and confirmation biases (Pariser 2011, Figá Talamanca and Arfini 2022). If the internet provides many interpretations of the same event, people tend to trust the source that provides the interpretation that does not force them to change their convictions. An additional, previously unencountered issue for open societies that promote (relatively free) public discourse is the installation of bots and ill-meaning agents promoting the divisions by spreading targeted misinformation. Some of the more disruptive election results of the 2010s are traceable to have been subject to those kinds of attacks (Mueller 2019), in which foreign agents attempt to create particular public opinion by littering digital spaces like Twitter with fake accounts spreading "opinions." Some of those may have not even been actual human beings, but chatbots.

These developments have, without doubt, changed the way many people interact with others in matters of politics. Even if those debates have not changed when they are performed in person, the digital public discourse has become a facilitator for fast-spreading, more pronounced opinions, group thinking and confirmation biases, as well as verbal aggression like threats and insults. The growing importance of the digitization for public discourse justifies a look into the social relationships build in what we call "digital society." Not only are people using the internet and its offers privately, but many areas of public discourse have moved from the press and other traditional media, social gatherings like clubs, or public speeches to digital platforms of social media. With people now

interacting with each other who would otherwise never have the chance to know of each other's existence, the term "digital society" appears more than adequate, and the concentration on its changes justified.

3.2.2 Private Relationships

The influence of how digitization has changed private social relationships does not only depend on the own use of social media. Unsurprisingly, if someone is not using those technologies, these new kinds of social relationships will remain absent in the immediate contact. However, the ubiquity of digitization suggests almost everyone has witnessed a change in some relationships. Reasons and causes for not using the new digitized infrastructure are manifold but often rest on a certain resistance to adopting new technology, or by simply not having those technologies available. Reasons for resisting to adopt new technologies are also diverse. Often, the lack of familiarity is a decisive factor for people foregoing participation. Many technological customs rest on a previous familiarity with other technology which may not be a given for every user, thereby forbidding any orientation with new technology interfaces and increasing the distance between user and technology. Thereby, without such familiarity, using technology, even when available, may be impossible (Vaportzis et al 2017). This also increases the perception of risk, as a lack of understanding and familiarity with technology is leading to overestimating risks associated with technology (Digmayer and Jacobs 2016). New kinds of social relationships may also be assessed as a risk to one's own lifestyle, as those are mediated through a technology that is wholly unfamiliar.

On the other side, those who do adopt and incorporate new technologies into their lives may find themselves establishing genuinely new forms of social relationships. Similar to how the postal system has made pen pals possible, digital means of communication like social media platforms have made genuinely new social connections possible. Those new connections not only do not fit the given social descriptors but are also changing established social conventions as well outside the new forms of technology. Thereby, even those who resist new social technologies may be affected by other people using them.

Not only the quality of given social connections may change while new types of relationships emerge, but also the quantity of social interactions has increased. Current internet users often are equipped with more

than just one way of being able to be contacted—email addresses, social media accounts with direct messaging function with either lasting or self-destructing messages, being tagged in people's posts "public" or "private" (i.e., a selected audience), and many more. Compared to the availability of people just a few decades ago—with only one phone in the house—the ease with which people can be reached (and with which they can present themselves to attract a different kind of people to relate to) can count as a fundamental change.

This increase in access not only works horizontally but also vertically, further mudding what is "public" and what is "private": entertainers (like singers and actors), politicians, important intellectual figures are often active on those social media platforms. Some aim to merely suggest accessibility, while others appear to interact with their fanbase, electorate, or general public online substantially. The ability to gain presence in the public discourse for almost any user has become an issue within the structures of debates, as social media can elevate even private pronouncements on the center stage of attention, if only for a few hours.

3.2.3 Embodiment and the (Un)Importance of Physical Presence

The developments in both the public as well as the private sphere show that many modern, digital relationships are not substantially mediated through physical presence. People fall in love with others without ever meeting them in person; psychotherapists are offering online-only sessions; coworkers may never meet while working cooperatively on a project (Chayka 2015). These forms of indubitably social relationships all have been made possible by the increased progress in digitization, but also by shifting social expectations of required physical presence.

It is crucial to acknowledge the decreased relevance of embodiment for many areas of social relationships (Lomanowska and Guitton 2016). Due to digitization, physical proximity and availability are no longer of the same relevance as they have been until this very moment in time. For one, because without physical proximity, establishing social relationships have been virtually impossible, except for sending letters or surrogates to do one's bidding. However, even those means always have been intended to lead to eventual physical closeness and interactions.

In contrast, many online relationships today are not intended to lead to being in the same physical place at the same time. Further, those who used the limited means available before the explosion of social media,

video chatting platforms, and the constant availability online were often pathologized in their behavior. Those who preferred talking on the phone with others without ever wanting to meet them in person were considered anti-social. In the same paradigm, people in the 90s and 2000s were diagnosed with "internet addiction"[1] due to their time spent in online communities, disregarding that these were early digital safe spaces for oppressed communities. Fittingly, the latest version of the Diagnostics and Statistical Manual of Mental Disorders (DSM-5) does not carry this diagnosis anymore (APA 2013).

Furthermore, with the widespread technology acceptance and increased potency of technological connections, the internet has become a vital part of many people's lives, not as an alternative to "reality," but as an additional layer to reality, with specific social relationships, norms, and customs.

3.2.4 Are "Facebook Friends" Friends?

In order to describe these processes and progresses on social media, the established terminology to describe social relationships has been imported and applied, often to the disagreement of those not participating in the digital social media. One leading example may be one of "Facebook friends." When Facebook started as one of the dominating social media platforms, their way of connecting people was by having users add each other to their respective "friends" list. Moreover, with differing uses, some people's "friends-list" grew to several thousand, while others intended to include only people that they would identify as "friends" offline as well.

Many cultural pessimists diagnosed a decrease in meaningful human–human connections due to the superficial nature of everybody calling each other "friends." While merely being acquaintances, coworkers, or random people one met at a party once (Mazie 2014). In order to regain

[1] Internet addiction has been split up to describe specific forms of the disorder, and one has been "Cyber Relationship addiction," which is supposed to denote the addiction of "using online relationships to replace real-life friends and family" (Young 1996), proving that this diagnosis was conceived from a strong norm-standard. This is not intended to relativize some obsessive uses of the internet which causes people to suffer. However, the generalized pathologization of "excessive" internet use, even though many internet uses are pro-social, interactive, and helpful, made it possible to abuse this diagnosis for merely educational purposes. Kershaw (2005).

the power over the distinction between "real" friends and those recognized as friends on Facebook, the prefix "Facebook-" was added: one has friends, and one has Facebook friends. The essentialism of who can or ought to count as "friend" aside, this distinction shows two things: on the one hand, digitization creates new social categories that previously have been non-existent. In general, technology often facilitates the enrichment of possible social relationships. This enrichment of the variety of social relationships is arguably one of the meanings of globalization—not only "grows the world smaller," but also the social network grows thicker. With people establishing more and new relationships with others all over the world, new technology may provide relational space for currently unthought-of relationships.

Without letters, there are no pen pals. And without digital social media, there are no "Facebook friends." Thereby, in denoting relationships of this kind as mere "Facebook friends," those cultural pessimists acknowledged that these are genuinely new relationships.

On the other hand, this new category allows for some disagreement while spelling out the underlying difference in uses of the chances of digitization: While the term "Facebook friend" denounces the superficiality of those relationships as "merely digital contacts," it also helps to recognize that many people have established friendships that are not "merely" but "purely" digital (see, e.g., Vallor 2012; Elder 2017; Kempt 2022). The ability to communicate daily allows friends from very different places in the world to remain involved in each other's lives. The social options differ in magnitudes to just a few decades ago, in which one's pool of friends was mostly limited to people from the same town.

And not only does this apply to friendships, but also to most other forms of social relationships (except for those predicated on physical proximity, like neighbors or roommates).[2] People have crushes, fall in love, spend time together, exchange sexual pictures and videos, and watch movies together—in short: they spend quality time together—without ever having to meet.

[2] Interestingly, there have been ways of relating to "number neighbors," in which people text the phone numbers right above or below their own phone number. (Ansari and Ries 2019).

3.3 Assigning Social Descriptors

With these exploratory questions set out, we can now turn the attention to characterize social relationships. As stated in the introductory chapter about the invested methods, language is not merely a tool of reflection and thinking. Since the linguistic turn in philosophy, we recognize that language is part of how we approach the world and our mind. (Glock and Kalhat 2018) That makes language fundamental to any form of analysis. However, after speech act theories (Austin 1962; Searle 1969) gave us an improved understanding that language is always happening in practical contexts, i.e., in contexts where we act, the pragmatics of language have become more important. This lead to a "pragmatic turn" within the linguistic turn. Therefore, the way we use descriptors to describe and characterize a social relationship will, in the end, influence how these relationships are being perceived and performed.

For example, the difference in social implications and internalized expectations of referring to someone as a friend or merely a co-worker is rather apparent. Social descriptors reference specific social protocols and conceptual ramifications, like the conceptual requirement of being familiar with a friend (otherwise one could not be friends in the first place and would use the term "friend" wrong).

The assignment of social descriptors only is descriptive because of their associated conceptual ramifications and conventional social protocols (Goffman 1956). Another example is the use of the word "friend." Through the conceptual dilution of the term "friend" on social media, where every contact is referred to as "friend," people started to refer to their friends that still fulfill stricter requirements as "actual friends." In contrast, the (somewhat presumptuous) reference to some contact on social media is referred to as "Facebook friend," as pointed out just above.

The action of describing social descriptors is collective, even though individuals perform the process of ascribing itself. Individuals have to choose from a collectively constructed set of social descriptors, which pre-structures possible social relationships of human beings. Those linguistically pre-structured social relationships are neither finite nor exhaustive. They are continuously changed, and some new ones emerge while some others fall into irrelevance. However, they are as "real" as other social constructs as well. Without descriptors like "husband" or "wife" (e.g., due to the lack of specific institutions, like marriage), those relationships

would not emerge. Of course, there could be marriage-like relationships, but with a different descriptor and possibly different associated expectations. Language does not just name certain activities and social relationships, but assigning a name to certain relationships allows for a sedimentation of habits and build-up of expectations and norms on how to behave in those particular social relationships.

It is crucial to notice that with this approach, for social relationships to be recognized, they must be assigned descriptors. Assigning those descriptors is not random but depending on the rules of assigning those descriptors. Those rules are indirectly negotiated through the language games of naming and describing: in being part of a group of language users, the continuous use of certain words for certain states of affairs sediments into the rules of how to use those words.

This reconstruction is very close to the ideas developed in Wittgenstein's "Philosophical Investigations" (Wittgenstein 2009, §43 and 138). The "use theory of meaning" Wittgenstein's claims that the meaning of certain words is determined by the way these words are used. This approach is an alternative theory to the common "reference theory of meaning" where a sentence or a word have their meaning because of their supporting reference in reality. A sentence "depicts" the world as it is and has meaning that way. A word has meaning because it refers to a specific thing in the world. Social relationships seem to have been understood in this referential way as well: the name for certain relationships is invented to describe certain social facts. However, in transferring the use theory of meaning onto the naming conventions of social relationship, we can understand the role of language in this better: through the naming of individual relationships and the sedimentation of those names (e.g., facilitated by certain technologies), the rules for identifying other relationships of this kind is made possible.

These conventions are often so familiar and self-evident that most people would not perceive them as anything but natural. Another example may help at this point to clarify. Nomad tribes that do not have a settled lifestyle with houses and streets may not have an elaborate term for "neighbor," since the specific location where one sets up their sleeping quarters is temporary and lacking relevance to reserve relational space for such a situation. They do not need a term for this situation, and consequently, any social protocol that may be attached to "neighbor" is empty. Now imagine, this nomad tribe decides to settle down and build houses for each member of the tribe. What has been fairly irrelevant until this

point, the physical proximity to other people, has become more relevant to them as they plan to reside in their places for the foreseeable future. With this practical relevance coming into the social setting of a group, a new term of a social relation appears adequate: the neighbor.

Many social descriptors, then, function to mark down an area in something we can call "relational space," which is the space according to which others can relate to the described person. They function as a heuristic according to which certain social protocols can be expected and enacted. Describing a relationship in a certain way communicates clues about the expectations toward the customs and norms of this relationship. Every new social relationship expands this relational space by adding to the manifold ways of how people can relate to each other. Thereby, the way in which people can relate to each other is unlimited, and an essentialist concept of social relationships seems inappropriate.

3.4 From Social Descriptors to Social Relationships

However, such a position renders most social descriptors connected to the relativity of cultural codification, since the contents of social descriptors vary with the cultural background. "Cultural codification" means that no characterization of a social relationship is free of already invested assumptions of the describer. Our tools in describing social relationships are shaped by the cultural conventions and uses of language, and may not fit in capturing the scope and depth of certain relationships. However, it seems that some social philosophers believe in the natural occurrences of certain social phenomena that can be described with certain essentialist conditions. Questions like Aristotle asked them about the "nature of friendship," then, are not only highly presumptuous about the way social relationships form, but also ignorant about the many varieties of such a supposed single term.

In his Nicomachean Ethics, Aristotle offers three different types of friendships, with different degrees of value (Aristoteles 1999). The lowest one is a "utility friendship," which is useful to one or both friends. The next is a "pleasure friendship" in which friendship is the source of pleasure for one or both. Lastly, a virtue friendship is supposed to be one in which mutual goodwill and well-wishing are the defining features, with a shared set of values. For Aristotle, the last one of the most valuable as

it is conducive to a virtuous and hence a good life, while the other two, though beneficial, are not leading toward a virtuous life.

Due to Aristotle's program of showing how friendships can contribute to a good, virtuous life, his friendship-reconstruction is strictly normative. Only friendships of a specified kind may add to an individual's virtue. The essentialist issue arises with the move that every friendship can be analyzed with these parameters, and philosophers use these parameters to approach and assess every social relationship of a specified kind.

In the context of the debate whether robots and humans can be friends, John Danaher (Danaher 2019) summarized the debate around robot friendships and accepted the challenge of adding four features of this highest form of friendship and asked whether a machine can fulfill these: mutuality, honesty, equality, and a diversity of interactions. It seems reasonable to expect those social relationships that we call "friendships" being built on mutuality, honesty, equality, and a certain diversity of interactive instances.

Danaher goes on to argue that most of these conditions are mere technical issues that with enough technological sophistication, we may overcome them. Some others, he argues, might be perceived as metaphysical impossibilities (ibid., 6f). However, it is questionable whether these are features of friendships per se, or whether these are features of good friendships. Are we mistaken in calling someone a friend if the relationship is not entirely mutual? It seems like we are not, as we may be somewhat disappointed but still consider the less interested party a friend of ours. Similarly, the fact that we talk to a friend only on a video messenger and never see them in person, which would violate the "diversity of interactions" requirement, seems to set a qualitative requirement of friendship, not a conceptual one. Thereby, Danaher follows many others in the debate about robot friendship by reducing philosophically interesting "friendships" down to "good" friendships here, not "friendship" as a social relationship itself.

Nevertheless, by defining the term "friendship" qualitatively, as many people have done following Aristotle, they suggest that there is an essential quality to friendships. Sven Nyholm (Nyholm 2020, 149) elaborates that many different versions of friendships are possible, even with machines, but the "highest form of friendship," (cf. Elder 2017) which is the virtue friendship, requires certain qualitative features that not every social relationship we call friendship fulfills. Namely, he worries that those four conditions of virtue friendship are required for fulfilling friendships.

Without being able to achieve these highest goals with robots, as Nyholm argues, other consequentialist concerns take over and should move us to reject the idea of robot friendships, like issues of alienation with other humans (ibid., 151; for a more elaborate analysis of this complex see Kempt 2021, 2022).

Moreover, while this is undoubtedly correct, it stands to debate whether this is a helpful way of both approaching social human–human relationships and social human-nonhuman relationships. First, because some explicit normative assumptions are transported with his terminology, and second, the silent premise here seems to be that friendships that theoretically cannot reach the highest level (due to metaphysical concerns, for example) are somewhat lacking. Yet, especially the latter is never elaborated upon, and it seems like it deserves some justification.

These considerations show that essentialism about social relationships, i.e., the thesis that certain social relationships have to exhibit certain unchangeable qualities, is inadequate. There are specific conceptual requirements to identify social relationships (as the standard example about conceptual truths "all bachelors are unmarried men" shows), without no naming convention would work. However, insisting on normative features of those relationships requires a rigorous interpretation of social relationships. And rigor rarely serves to praise certain relationships and mostly serves to dismiss others.

The rigidity, then, is the primary normative issue with this kind of social essentialism: philosophers and laypeople alike have used this theory to devalue and dismiss many relationships. Essentialism like this allows disregarding people in those relationships who claim that they are just as valid as those falling under the essentialist definition. The claim that marriage can only be between a man and a woman dismisses the notion of same-sex love, and to this day this essentialism justifies anyone who puts into question the quality of romantic relationships of many kinds, like the love across races, classes, ages, abilities, and religions. In a similar vein, people have proposed that women lack the virtue to enter those high-quality friendships, thereby cementing the patriarchal structure of ruling men and ruled women. See for example Plato's thesis of the "inferiority of the soul of women" in Timaeus 42a (Plato 2000).

Essentialism proposes a possibly biased, hard definition of what a particular social relationship is or could be, before taking into consideration the experiences of people in those relationships. A use theory of social relationships proposes to understand relationships the other way around:

with only a light conceptual framework at hand, assigning descriptors to social relationships are taken from the performance of those relationships. The understanding that the quality of certain social relationships, especially the notion of "true relationships" can be defined at all ought to be dismissed.

The consequence of an anti-essentialist position on social relationships is that there are no "true friendships" or "true romances." Evidently, that does not affect the quality of certain relationships. Just because we reject the idea that once all requirements are fulfilled, two people have achieved a "true friendship," people will not like each other less or treat each other with less dedication. We may even call those friendships and romances exemplary or worth following if we realize that there is no true ideal to follow but the conviction of two people to be in a satisfying social relationship.

3.5 Agent-Network Theory

The proposed relational approach has several intellectual neighbors that deserve mentioning here but cannot all be addressed. One of them is the so-called agent-network theory (ANT) which seeks to understand the social as a collection of relations as well. Bruno Latour (Latour 2004, 2005), one of the leading proponents of this approach, developed ANT as an alternative to a methodological mistake he claims sociology has made: In not reflecting upon the concept of the "social," but defining it, other sociological theories presuppose what they claim to investigate. In the disguise of explaining the social, then, they are simultaneously constituting it. ANT allows for de-centering the subject, as it is understood as embedded in a network of other agents and in relationships with them. ANT is not limited to describing social relationships alone but is a methodological approach to describe phenomena via the analysis of relationships, e.g., the relationship in observing and interpreting animal behavior through analysis of the relationships of those involved in the interpretation of the animal behavior.

In a digital society with an immensely increased connectivity and thereby social relationships, such an approach provides the method to understand social developments through social relationships.

The methodological claims in ANT, however, are fairly strong, as they reject any explanation of the social that does not start with describing relationships. As stated at the beginning of this chapter, the approach

here is not to unsettle social theory by criticizing its foundational assumptions, as Latour aims to. The relational approach here does not share those substantial methodological criticisms of sociology as a science but rather seeks to establish the undetermined and principally open nature of the social as a relational phenomenon. The social, in the view purported here, can reasonably be applied to more than merely human–human relationships, as it is constructed by social descriptors and scripts rather than specific features of members of society.

This approach does not contradict the standard reconstruction of the social as the study of human society. However, it may add the asterisk that "human society" only has been the default due to a lack of alternatives. Additionally, we can argue that with the dawn of communicative AI, such an alternative is now available. However, ANT and other social network theories are important theories in the effort to recognize the relevance of presuppositions and preconceived notions in describing social relationships, as they also lay the groundwork for some ethical theories on moral patiency (Gunkel 2012, 2023). Lastly, while moral patiency is a topic of discussion in another chapter, the very idea that not everything in the concept of the "social" has to be an equal social agent can be established here.

3.6 Gender as Relational Descriptor

In an even greater essentialist assumption throughout the cultural history of humanity (Bem 1993), gender-differences have been taken to be associated with biological differences of the sexes and thereby predetermined as a biological determination of how society ought to be organized. These biological differences were invoked as justification for patriarchal power structures from primogeniture to the exclusion of women to hold public or religious offices to economic dependencies. Since its earliest movement, many people criticizing the social protocols and expectations associated with a certain gender-description as based confusing gender and sex. Simone de Beauvoir summarized the early issues by stating that "one is not born, but, rather, becomes, woman" (De Beauvoir 1949, 330) pointing toward the performativity of gender based on certain gender roles and expectations.

Through the introduction of the distinction between gender and sex, many social protocols turned out to be based primarily on gender (Butler 1988). The realization that many of those norms and protocols are based

on gender, however, makes them circular: if gender is a construct, one cannot deduce social roles and certain social descriptors based on gender, as the construction of gender is determined through the descriptors of the relationships between genders.

Gender has been one of the key social descriptors along which certain relations have been formed, generating a particular heuristic according to which people can form their expectations. Identifying someone as a woman or a man provides some information, as the social descriptors of gender are associated with certain protocols, but more importantly a certain "familiarity" with the other person.

The assumptions about the naturalness of those protocols due to the familiarity with them are another instance of essentialism, as any reference to "necessary" or conceptual elements of gender-specific social descriptors presupposes unchangeable features. Obviously, this description is limited to human–human relationships, even though it seems to have a strong pull for engineers and some AI theorists as a social heuristic worth keeping. Some argue that since "female voices lead to users assessing their device as more trustworthy" (Steele 2018), the replication of the positive features of gender stereotypes ought to be used to improve upon the device's efficiency with the user. Clearly, the overall principle here is that anthropomorphism is a helpful metric to create relatable robots. However, it is clear that this approach rests on the essentialist assumptions of gender and further establishes the certain gender- "typical" features as given and unchangeable.

For the suggestions developed in the next chapters about expanding the relational space to include human–machine relationships, it is worth keeping in mind that one of the main social categories, gender, is best understood as a social construct, not a biological fact.

3.7 Institutionalized Relationships and Relationships qua Humanity

Having defined relationships as mutually consequential, one could ask whether people stand in some relationship with other humans by merely being human. For example, whether someone is in a social relationship with a coma patient whom they have never met before is unclear, since the coma patient will not react to anything they do, nor will they be able to form any bond. However, we might want to claim that there is a human–human relationship dimension that comes into play via humans being

humans. Shared humanity is usually invoked when insisting on describing even the most unconnected humans as standing in some relationship with each other. Coming from some religious traditions (like the idea that "we are all God's children") or some secularized approaches like human rights, the idea that there is a "shared humanity" that puts us all in a relationship with each other is well established.

Thereby, a coma patient is, despite their inability to have any meaningful interaction with us, still in a human–human relationship with us. This seems to be the constitutive core idea of humanists when arguing that neither animals nor artificial intelligent agents can be in a human–human-level relationship with humans: they lack the core feature of humanity, thereby limiting the relationships of the latter to humans to transactional or pragmatic relationships, not relations-in-principle. Constructing humanity as a fundamental feature to distinguish humans from animals and other entities does not downgrade or dismiss animals (it is thereby not speciesist). It is instead a concept to establish the idea of shared humanity that nobody can avoid, even as a hermit with no social connections whatsoever. From the social relation as "human qua human," humanist philosophers conclude that we owe each other some basic virtues.

Many modern institutions of civilizations are built on the idea that all humans are in equal relationships, as the rule of law or democracy shows. These institutions can, in this way, be understood as the realization of fundamental human–human-relationships (see, as a realization of this idea in the area of justice, the approach of "relational egalitarianism" (Young 1990; Heilinger 2021). This is not necessarily the most reliable way to establish the concept of humanity and the political theory that lead to the modern achievements of civilization. A relational approach to humanity is merely one way to provide a clear path to see society as analyzable through social relationships, i.e., the social descriptors we choose to apply to each other.

For a relational approach and the idea that social relations are not limited to humans only, this claim does not hold much of a consequence. The fact that humans stand in a privileged relationship with each other does not affect the ability to expand the social fabric beyond humans.

3.8 Domestication as a Social Technique

We stated before that technological progress often implies the opening of some relational space. Without a postal system, the social relationship of pen pals would be impossible. Similarly, without social media, the relation of being "Facebook friends" would not be possible. There is, however, an even more important technology, or possibly rather a technique, which has made some social relationships possible that refute the idea that meaningful relationships are only possible between humans.

Taming and domestication are cultural techniques to control animals for human use and companionship (Zeder 2015). With its invention several thousand years ago, the purpose for having animals domesticized (i.e., in their homes) was most often for immediate use: more effective hunting strategies (with dogs), keeping unwanted pests out of the house and food storage (with cats) (Ottoni et al 2017), easier agricultural food production (with horse-powered plows), warfare (with horses), and as a source of food and clothing (with sheep, cows, goats, pigs).

Moreover, while some of those initial purposes are still valid today, e.g., food production, some other uses have been replaced or fallen out of practice altogether (like the use of cows for plows or horses for warfare). One purpose, however, has been constant in many different cultures and is today the predominant use of individuals for domesticized animals—they are pets. A pet, i.e., a domesticized animal with no immediate purpose, is most often kept purely as company. Today, people bring cats in their homes not because they have a rat-problem, but because they appreciate another independent being in their homes.

Next to the relationship-enabling features of domestication, a field of theories in technology studies has evolved under the idea of a "domestication theory of technology" (Berker et al. 2005). This approach describes the processes with which users of a technology appropriate and "tame" a given technology in their own purposes, as humanity has done thousands of years ago with wild animals. The idea of "taming" technology is more akin to subjecting it to one's own specific purposes in domestic settings (Silverstone and Haddon 1996). In this view, technological innovation is often created by the appropriation of a given technology to new contexts, through domestication (ibid.). This view can help us appreciate the transformative power of technology in unforeseen ways once it is "in actual use," although our focus lies on the idea that domestication as a technique is first and foremost a relational technique.

The reason why the technique of domestication is so relevant for the concept of the social consists in the fact that people establish unique social relationships with the domesticated subject—be it an animal or some technology. Additionally, as stated before, without this technique, these kinds of social relationships were not possible: Pets are mostly kept as company. For most, they do not replace human relationships but are incorporated as an additional layer of the relational network. These relationships are in almost all cases not comparable to human–human relationships (some exceptions may apply to tamed great apes) but constitute their own category. However, with the distinctions laid out above, these still count as social relationships as they have consequences for both parties involved and feature certain embeddedness in their overall social network.

In the context of the essentialist debate, human-pet relationships demonstrate the mistaken assumptions of such a position: if we asked essentialists prior to any domestication or taming effort whether relationships to animals other than hunter/hunted would be possible, they probably would have rejected the idea of human-nonhuman relationships. An essentialist can only define the relationships that are, not those that may. The pragmatic approach, however, recognizes that anything capable of entering a consequential, embedded relationship with others is, in principle, a possible addition to the social fabric.

One may interject that those relationships, if perceived as especially meaningful, are the result of inadequate and mistaken projections. Some people do, in fact, treat their pets like their children or friends, applying psychological explanations for the pets' behavior that are more befitting the behavior of a small child. Any theory that cannot explain those projections as mistaken about the fundamental depth of the relationship opens up issues with arbitrariness. Since pets cannot give an account of their perception of the relationship, humans tend to read whatever they want into those relationships. Fortunately, the approach proposed here does not encounter these issues. The main point here is especially that human-animal relationships are not describable with human–human relationship terminology, rendering those anthropomorphic transfers problematic and almost always wrong. However, this does not mean that those relationships in themselves cannot be social or meaningful.

The very idea proposed here is that this is a different category of relationship but a relationship nonetheless, and that a particular technique—the technique of domestication of animals—has made this kind of relationship possible. Describing and measuring human-pet relationships against human–human relationships must fail, as those relationships are of categorically different kinds. Nevertheless, this categorical difference does not render them invalid or meaningless. Essentialists who argue against human-pet friendships are not mistaken about the quality of human-pet friendships; they are mistaken about applying their assumptions about "friendship" to human-pet relationships. In short: the category of "social relationship" is not exclusively human.

While technological progress like the letter, social media, and others have made certain human–human relationships possible, the importance of domestication lies in the fact that it made human-animal relationships possible for the first time.[3] And this transference, as we argue later on, also applies to artificial intelligence and its associate autonomous agents. Thereby, AI should be considered less on the level of an invention or discovery of a particular technology, like fire (Google's CEO Pichai claimed AI to be as impactful as fire or electricity, [Goode 2018]), but more on the civilization-making technique of domestication.

3.9 Against Relational Arbitrariness

It seems appropriate to say a few more words on one main worry about rejecting essentialism about social relationships: that with the rejection of essentialist guiding rails of social relationships, we are headed for social arbitrariness. The main worry here is that the pragmatic approach of social relationships runs into issues with some counterintuitive claims of social relationships. Since a pragmatic approach does not rely on hard definitions of "true relationships" against which the contents of individual relationships are measured, people may claim relationships with objects and other entities, like with holograms (Gollayan 2018). The issue here is mainly relationships with supernatural beings, para-social relationships, and relationships with aliens.

[3] Needless to say, that domestication as a game-changing technique of civilization has also made a variety of new human–human relationships possible due to the relief of work, the availability of food, bigger coverage of distances, and so on.

3.9.1 Supernatural Relationships

Supernatural beings have been part of the human social fabric metaphorically and metaphysically. Metaphorically, "spirits" of deceased members are sometimes still considered to be among the living, with certain influences and obligations of the living toward the deceased. Piety for the dead is not only a metaphorical way for the living to generate norms that guarantee traditions and the stability of social interactions but has long been counted as an element of social interactions with the dead. However, since these are purely metaphorical and beyond any reasonable evidence, those interactions are not considered social, as these relationships are not consequential in both directions, and even less reciprocal.

Similarly, some people claim to be speaking to God(s) directly, which would constitute a supernatural social relationship (this is explicated in the concept of a "personal God" [Wainwright 2012]). Many societies have the concept of social human-God relationships, with differing degrees of accessibility. Prophets, priests and priestesses, mediums, and rainmakers are usually considered those who can stand in direct contact with some supernatural entities. These (projected) relationships of humans to some supernatural influences have been constitutive of some basic human institutions like morality, theology, and some forms of governance (like law and constitutional principles).

If there are humans with consequential access to Gods, and many cultures developed the idea that it was at least theoretically possible for humans to stand in a certain privileged human-God(s) relationships (like prophets), then this would constitute a genuinely new social category of relationships.

However, the metaphysical and epistemological issues encountered with claims of being in a human-God social relationship are relevant here. Those who do not believe in supernatural entities will have a hard time understanding and accepting the contents of these supposed relationships. The main philosophical issue, however, is one of epistemology: without being able to share these relationships with others, they remain inherently asocial. Relationships with the supernatural are not embedded in a social-relational network but are discrete. One of the few conceptual requirements we proposed for social relationships at the beginning of this chapter was that they could be embedded, even if people keep them secret. They need to be demonstratable, and human-God relationships are

hardly so, even if the presumption of the possibility of those relationships clearly affects the way cultures are structured.

3.9.2 Para-Social Relationships

The phenomenon of the para-social is related to the supernatural, and human-God relationships are often portrayed as paradigmatic "para-social relationships," as the one-sidedness of people's relationships with supernatural entities presents itself as if it was a fully two-sided relationship. However, it seems worthwhile to differentiate the supernatural from the worldly para-social phenomena because of differences in social relevance (in which the former is outweighing the latter) and the relatability and accessibility (in which the latter is more important than the former).

Para-social relationships are one-sided relationships, where one person extends emotional energy, interest, and time, and the other party, the persona, is completely unaware of the other's existence. (Horton and Wohl 1956) Some of the most common candidates to be the persona, i.e., the entity onto which people project their relationship-illusion, are celebrities and fictional characters. (Liebers and Schramm 2019) Of course, not every fan of a particular pop artist is in a para-social relationship, even though this term has been used to describe the impact of media characters on a consumer; as stated above, for the relationship between two entities to be considered social, it must be consequential for both parties, even if not equally consequential. The relationship between a fan and a celebrity is, overall, consequential for both parties, as the celebrity without fans would not be famous. Para-social relationships with celebrities are, in this regard, suffering from misattributions about the correct social descriptor and thereby the character of their relationship.

Moreover, while mischaracterizations of the extent and intensity of relationships could be considered para-social, as some people are mistaken about their relationships, some other forms of para-social relationships are categorical mistakes. People claiming to be in a relationship with fictional characters like James Bond (ibid.) are indubitably mistaken about the attributability of relationship descriptors to fictional characters. And in this misattribution, they could be seen as mistaken about the concept of relationships, as fictional characters do not behave in a sense that fits the same category as human behavior. Even more problematic are para-social relationships with organizations, even though those are often metaphorical or not claimed to be "relationships" in any comparable sense to social

relationships. Fans of some sports teams may claim to be "married to their club," but this does not mean that they see themselves as unavailable for other romantic relationships. In contrast, those claiming to be in an intimate relationship with their favorite fictional character often are similarly infatuated as people in love.

Para-social relationships, then, do not have to be considered moving forward since they do not meet our definition of social relationships. Thereby, pragmatic theories of social relationships have tools to resist the accusation of being too allowing in the reconstruction of the concept of the "social."

3.9.3 Relationships with Aliens

It may seem presumptuous to spend any time on a topic currently best left to science fiction writers. However, some small remarks on how we may relate to alien intelligences are appropriate here, considering that mathematical models and astronomic discoveries over the past decades suggest that we are not alone in the universe (Shostak 2018).

The question of how we would (or will) relate to alien intelligent agents is more informative if understood as a thought experiment: the strategies deployed to answer the question of how we could relate to those entities tell us more about the concept of relating invested in those strategies. One main observation here is that nobody would suggest approaching those agents with trying to establish relationships akin to human–human relationships. It would count as a categorical error to assume that their way of relating to each other is anywhere close to how we relate to each other. Asking what a "human-alien friendship" would look like and if it would ever reach the Aristotelian level of virtue friendship demonstrates the essentialist assumptions quite nicely.

We can safely assume that we would immediately and self-evidently expand the relational space to cover all possible human-alien relationships that can be established with them. Thereby, the flexibility of the social fabric, i.e., the willingness of people to include all sorts of social relationships, from non-present human–human relationships to human-pet relationships and possible human-alien contacts, is a necessary premise. Any concept of social relationships that determines which relationships are valid and which are not appears to be investing unjustified normative premises.

3.10 Lessons for Imminent Changes or: The Rise of Human–Machine Relationships

To summarize the approach taken here: social relationships are an important entity in analyzing the dynamics of social interactions. Taking this position does not intend to provide an elaborate sociological terminology, but rather a method with which the diversity of human connections, both with other humans as well as with non-humans, can be analyzed. It showed that technology changes the relational space quantitively by making people more accessible and providing genuinely new human–human relationships. However, technology also changed social relationships qualitatively and in two ways: first, some social relationships do not require physical presence of interactions anymore, as was previously required. New communicative technologies allow for purely digital social relationships to occur that are perceived by those in them as equally fulfilling as social relationships with people in their physical proximity. And second, since the domestication of animals, human-animal relationships can count as some form of social relationships due to their interactive, embedded nature and the relevance of those connections for both pets and humans.

Therefore, even categorical changes in social relationships are common and constant, and most people may not even realize them. Humans are fundamentally social animals that are capable of relating not only to others in their immediate proximity but also—somewhat abstract—to the entirety of humans via their own humanness.

With these preliminary inquisitions in place, the realization that qualitative and quantitative changes in social relationships often depend on technological progress allows for a turn toward the current technological revolution: artificial intelligence. We observe that people start to build relationships with artificially intelligent agents that may count as social relationships, even as relatively primitive ones. There is noticeable reciprocity in those relationships, even though this reciprocity is not derived from a state of mind (like emotions), but pragmatically. Due to a lack of terminology and established social human–machine relationships, the description of those relationships resorts to anthropomorphic projections and terminology.

And because of those anthropomorphic projections, a discourse developed that questions the very possibility of human–machine relationships.

People who are willing to build those relationships are often being ostracized by a critical public, who read into the ever-growing number of AI appliances a swift, uncalled for, and potentially dangerous shift in the social fabric and human's relationship with technology. And while it seems unclear where these debates go, the two motivating issues for this project will remain: people will keep building relationships with these artificial agents, and the sophistication of those agents will increase as well. Both of these together justify a closer look into the underlying social descriptors and may warrant an expansion of the relational space into social human–machine relationships.

References

American Psychological Association, APA (2013). *DSM-5*. Washington DC: APA Publishing.

Ansari, Zoya, and Ries, Brian (2019). People are Texting Phone Numbers Identical to their Own, with One Key Difference. *CNN*. https://edition.cnn.com/2019/08/05/us/twitter-number-neighbor-trnd/index.html (Last accessed on August 27, 2024).

Aristoteles (1999). *Nicomachean Ethics*, translated by Terence Irwin. Indianapolis: Hackett

Austin, John (1962). *How to Do Things with Words*. Cambridge: Harvard University Press.

Bem, Sandra (1993). *The Lenses of Gender: Transforming the Debate on Sexual Inequality*. New Haven, Connecticut: Yale University Press.

Berker, T., Hartmann, M., and Punie, Y. (2005). *Domestication of Media and Technology*. McGraw-Hill Education (UK).

Bouvier, Gwen, ed. (2018). *Discourse and Social Media*. London: Routledge.

Butler, Judith (1988, December). "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." *Theatre Journal. Johns Hopkins University Press* 40 (4), 519–531. https://doi.org/10.2307/3207893.

Chayka, Kyle (2015). Let's Really be Friends. A Defense of Online Intimacy. *The New Republic*. https://newrepublic.com/article/121183/your-internet-friends-are-real-defense-online-intimacy (Last accessed on August 27, 2024).

Danaher, John (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies* 3(1), 5-24

De Beauvoir, Simone (1949). *Le deuxième sexe* (The Second Sex). NRF Essais.

Digmayer, Claas, and Jakobs, Eva-Maria. (2016). *Risk Perception of Complex Technology Innovations: Perspectives of Experts and Laymen*. https://doi.org/10.1109/IPCC.2016.7740510.

Durkheim, Emile (1985). *Les Règles de la Méthode Sociologique* (The Rules of Sociological Method). Paris: Felix Alcan.

Elder, A. M. (2017). *Friendship, Robots, and Social Media: False Friends and Second Selves*. Routledge.

Figà Talamanca, G., Arfini S. (2022). Through the Newsfeed Glass: Rethinking Filter Bubbles and Echo Chambers. *Philos Technol.* 35(1), 20. https://doi.org/10.1007/s13347-021-00494-z

Glock, Hans-Johann, and Kalhat, Javier (2018). *Linguistic Turn. In: Routledge Encyclopedia of Philosophy*. London: Routledge.

Goffman, Erving (1956). *The Presentation of Self in Everyday Life*. New York (NY): Doubleday.

Gollayan, Christian (2018). I Married My 16-year-old Hologram Because She Can't Cheat or Age. *The New York Post*. https://nypost.com/2018/11/13/i-married-my-16-year-old-hologram-because-she-cant-cheat-or-age/ (Last accessed on August 27, 2024).

Goode, Lauren (2018). Google CEO Sundar Pichai Compares Impact of AI to Electricity and Fire. *The Verge*. https://www.theverge.com/2018/1/19/16911354/google-ceo-sundar-pichai-ai-artificial-intelligence-fire-electricity-jobs-cancer (Last accessed on August 27, 2024).

Granovetter, Mark (1985). Economic Action and Social Structure: The Problem of Embeddedness. *The American Journal of Sociology* 91(3), 481-510.

Gunkel, David (2012): *The Machine Question: Critical Perspective on AI, Robots, and Ethics* Cambridge, MA: the MIT Press.

Gunkel, David (2023). *Person, Thing, Robot. A Moral and Legal Ontology for the 21st Century*. MIT press.

Heilinger, J-C. (2021). Individual Responsibility and Global Structural Injustice. Towards an Ethos of Cosmopolitan Responsibility. *Journal of Social Philosophy* 52(2), 185–200.

Horton, Donald and Wohl, Richard R. (1956). Mass Communication and Parasocial Interaction: Observations on Intimacy at a Distance. *Psychiatry* 19, 215-29.

Kempt, H. (2021). Zwischenmenschlichkeit für Maschinen. *KI–Die Große Verheißung*, 453–470.

Kempt, H. (2022). *Synthetic Friends. A Philosophy of Human-machine Friendship*. Palgrave.

Kershaw, S. (2005). *Hooked on the web. Help is on the way*. New York Times. https://www.nytimes.com/2005/12/01/fashion/thursdaystyles/hooked-on-the-web-help-is-on-the-way.html

Latour, Bruno (2004). *Politics of Nature: How to Bring the Sciences into Democracy*, translated by, C. Porter. Cambridge, MA: Harvard University Press.

Latour, Bruno (2005). *Reassembling the Social: An Introduction to Actor-network Theory*. Oxford/New York: Oxford University Press.
Liebers, Nicole and Schramm, Holger (2019). Parasocial Interactions and Relationships with Media Characters-An Inventory of 60 Years of Research. *Communication Research Trends* 38(2), 4–31.
Lomanowska, Anna, and Guitton, Matthieu (2016): Online Intimacy and Well-Being in the Digital Age. *Internet Interventions* 4, 138-144.
Mazie, Steve (2014): Do You Have Too Many Facebook Friends? *BigThink*. https://bigthink.com/praxis/do-you-have-too-many-facebook-friends (Last accessed on August 27, 2024).
Mueller, Robert S. (2019). *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. Department of Justice, United States of America. https://www.justice.gov/storage/report.pdf (Last accessed on August 27, 2024)
Nyholm, Sven (2020). *Humans and Robots. Ethics, Agency, and Anthropomorphism*. Lanham (MD): Rowman-Littlefield.
Ottoni, C., Van Neer, W., De Cupere, B. et al. (2017). The Palaeogenetics of Cat Dispersal in the Ancient World. *Nat Ecol Evol* 1, 0139.
Pariser, Eli (2011). *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. London: Penguin.
Plato (2000). *Timaeus*, translated by Donald J. Zeyl. Indianapolis: Hackett.
Searle, John (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.
Shostak, Seth (2018). *Simple Math Shows How Many Space Aliens May Be Out There*. SETI Institute. https://www.seti.org/simple-math-shows-how-many-space-aliens-may-be-out-there (Last accessed on August 27, 2024)
Silverstone, R., and Haddon, L. (1996). *Design and the Domestication of Information and Communication Technologies: Technical Change and Everyday Life*.
Steele, Chandra (2018). *The Real Reason Voice Assistants Are Female (and Why it Matters)*. Medium. https://medium.com/pcmag-access/the-real-reason-voice-assistants-are-female-and-why-it-matters-e99c67b93bde (Last accessed on August 27, 2024)
Vallor, S. (2012). Flourishing on Facebook: Virtue Friendship and New Social Media. *Ethics and Information Technology* 14, 185–199.
Vaportzis, Eleftheria et al. (2017). Older Adults Perceptions of Technology and Barriers to Interacting with Tablet Computers: A Focus Group Study. *Frontiers in Psychology* 8, 1687. https://doi.org/10.3389/fpsyg.2017.01687
Wainwright, William (2012). *Concepts of God, Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/concepts-god/ (Last accessed on August 27, 2024).

White, Cindy (2015). Reciprocity/Compensation in Social Interaction. In *The International Encyclopedia of Interpersonal Communication*.

Wittgenstein, Ludwig (2009). *Philosophical Investigations*. Oxford: Wiley Blackwell

Young, I. M. (1990). *Justice and the Politics of Difference*. Princeton University Press.

Young, K. S. (1996). Psychology of computer use. Addictive use of the Internet: a case that breaks the stereotype. *Psycholical Reports*, 79, 899–902.

Zeder, Melinda A. (2015). Core Questions in Domestication Research. *Proceedings of the National Academy of Sciences of the United States of America* 112(11), 3191-3198. https://doi.org/10.1073/pnas.1501711112

CHAPTER 4

The Basics of Communicative AI

4.1 Definitions

Definitions of what AI is and what it could be have been plaguing the debate about it from its inception to this day. Whether it is in legislative efforts, philosophical debates, or engineering departments, consent about a definition of artificial intelligence is notoriously hard to come by. Some argue that AI ought not to be defined at all, as any definition will limit the way we can interpret computational intelligence and possibly pose a hindrance in creatively approaching AI.

However, we are justified in using the definition of AI that supports the effort undertaken here, which is to explore how our relations with chatbots will develop. This definition is not meant to cover all possible versions of AI, nor does it aim to unify the debate to a coherent total. In fact, the opposite is the case: the definitory work undertaken here ought to be considered a small subsection of artificial intelligence. Thereby, it could be understood less as a definition, but rather as a semantic characterization of a specific technology. Primarily due to the wide variety of applications that operate with algorithms considered "artificially intelligent," such a unifying definition is neither required nor would one be helpful.

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_4

A constructive approximation of what can be considered artificial intelligence is laid out by Jacob Turner in "Robot Rules" (Turner 2019, 16):

> Artificial Intelligence Is the Ability of a Non-natural Entity to Make Choices by an Evaluative Process

Due to his focus on the governance and legal ramifications of AI technology, this definition focuses on two main functions of AI: the autonomous activity of such technology (by making choices) and the evaluative process that precedes those choices. How exactly the "evaluative process" unfolds is, as a result of this, not relevant, as it is to be understood as primarily an issue within engineering paradigms and the application of the AI. A sorting algorithm trying to match a customer with a product they most likely are interested in is going through an evaluative process as much as an autonomous car is when it navigates traffic. There is, however, a danger that ought to be kept in mind when using this phrase. Engineers may understand such an evaluative process as "assigning values to different options and choosing based on the most valuable option," with "valuable" being relative to the goal (i.e., the likelihood of a match, or highest outcome, etc.). Sometimes, this process is thought to be a representative model of how humans ought to make their decisions as well. That is, however, usually not how normative deliberations on how to act are coming about, as deliberative processes are often more complex and not quantifiable in a binary pro- and cons-list. Thereby, the term "evaluative process" is understood as the simple assignment of value-parameters under a predetermined goal.

This "evaluative process," then, is being complicated when other parameters ought to be considered. The ethical considerations are concerned with the incompleteness of the parameters that are involved in such an evaluative process, with some questioning whether those parameters can be broken down into computable pieces (i.e., whether machines can make moral decisions that are not hard-coded into their evaluative process). This is not limited to self-driving cars or killer drones but begins with matching algorithms, e.g., a search algorithm that can lie (Bendel 2018) or that matches suicidal requests with unquestioned answers to achieve that purpose.

How exactly this necessary second level of evaluation, the evaluation of the determined choice according to some ethical considerations, can be

computed is one of the main challenges for "moral machines" and subject to controversies surrounding AI ethics. However, the definition at hand is still helpful, as it can be expected that the second level evaluation will still be an "evaluative process," whether it be a utilitarian calculation, some pre-installed virtues, or the sorting of the choice under some action-types and the ability of their maxims to count as universalized rules. Therefore, "choices based on evaluative processes" should be considered a reliable approximation.

One dimension, usually considered part of AI, is missing from this definition. Knowledge representation (KR) is a concept describing the ability to use received information for inferences. As the basis for an AI's "understanding" of the world, this is the source from which the evaluative process operates (Brachman and Levesque 1985). Without knowledge representation, there simply would be no inferences, i.e., evaluative processes.

Since Turner is interested in the pragmatic dimensions of AI, as his goal is to provide principles according to which some normative frameworks to regulate AI can be built, knowledge representation is less of an issue for his definition. However, KR is considered the core methodological issue where many other issues with AI will be decided (Smith 1985).

4.1.1 An Additional Distinction

So far, AI is achieved in its most basic terms when it can make decisions based on an evaluative process. The confusion about finding an all-encompassing definition of AI might stem from the diverse range with which AI is being utilized. From self-driving cars to sorting algorithms to chatbots, every autonomous technology claims to be powered by some AI, mudding the definitory waters by confusing principles with concrete entities. Just because there are different uses of AI does not mean that AI itself is incoherent and, thereby, indefinable.

Yet, there are some useful internal differentiations to be made for different practical purposes. From a philosophical perspective, machines that utilize natural language to communicate with human beings may be of a different consequence than machines that merely behave based on gestures or in coordination with humans (e.g., sorting algorithms). This difference can be demonstrated when these developments are viewed from their (not yet achieved) practical end: a multi-purpose robot that can master every behavioral task (except for speaking) is of a different

philosophical relevance than a chatbot that can master every conversation. Even more significant is the difference between a highly sophisticated matching algorithm, like a music-matching service and a chatbot.[1] The former cannot interact with us directly, as the input is merely based on our listening behavior, possibly without the awareness that an algorithm is using this as data for providing more music. At the same time, the latter, through the use of natural language alone, provides a richer, interactive experience.

These observations suggest that there is a philosophically relevant internal distinction between different applications of AI, which will be called "conversational AI," in opposition to "behavioral AI." Phenomenologically speaking, being confronted with an autonomous machine that moves without any guidance through an obstacle course appears like an impressive feat of technology we can get used to as a new standard of what technology can do. However, being confronted with a machine that keeps a conversation with us on a sophisticated level has an entirely different phenomenological impact, as it not only walks the same space as we humans do, it also talks in the same categories as we do.

This does not mean, as stated before, that this is a technological difference or one that conceptualizes one to be more "intelligent" or more useful than the other, as the measurement for "intelligence" is even more opaque than a definition of AI. This distinction is to focus the attention on machines that converse with human beings using natural languages, while not discarding the others. One of the main applications of "conversational AI" is the chatbot, which are programs designed to respond to natural-language input with adequate answers in natural language as well.

What exactly distinguishes conversational AI from other forms of AI? We can define conversational AI within our definition as the forms of AI that make evaluative decisions based on natural-language input and produce natural-language outputs in return that are responsive to the input. Or, quite simply put, it talks. In opposition to other natural-language-using AI, like translation algorithms or voice-operated machines, conversational AI produces semantically meaningful sentences autonomously in a conversational setting. The terminology for chatbots that can keep conversations on virtually any topic are "open-domain" or "general-purpose"-chatbots, which will be used interchangeably.

[1] From a technical perspective, some chatbots may also count as matching algorithms that merely match inputs and outputs of natural-language conversations.

Chatbots fulfill this definition of conversational AI and are, in the standard tropes of Sci-Fi literature, the most likely candidates to cause issues that are not limited to ethical considerations but to questions in the philosophy of mind. For some reason, framing philosophical questions around AI presupposes an NLP-portion, as if it was self-evident that AI would be using natural language to communicate with us. Moreover, the main issues arising in the philosophy of mind debates about self-awareness and consciousness of those machines are predicated on those machines being able to communicate their mental states to us, using natural language.

This marks natural-language processing as a philosophically required form of AI or instead as an AI-premise for other relevant issues within the philosophy of AI. Thereby, it is justified to take NLP as a specifically relevant form of AI—not only because of its phenomenological impact but also because other issues within philosophy of AI are predicated on it.

Any AI that can establish and hold a conversation with human language users, then, is included here. Obviously, this opens up questions on when "holding a conversation" is fulfilled, and what level of sophistication a chatbot has to exhibit for the phenomenological impact to appear. And while these are pragmatic questions that will ultimately be decided by the available technology, the assumption that eventually, this technology will have that effect is justified.

While most state-of-the-art chatbots today pass the Turing test, but do not appear all that sophisticated once one knows what to look for, as they can be tangled in conversational loops fairly easily, misunderstand or do not understand at all phrases we expect them to understand. The advancements in the area of natural-language processing in the last few years, however, have produced chatbots that are, in fact, of a certain sophistication (see for details Chapter 8 on LLM). This development is largely due to advancements in the training of these machines.

4.1.2 Embodiment

The most advanced chatbots are text-based, while standard personal assistants are voice-operated (like GPT-4o, Alexa, Siri, Gemini, or Cortana). Both are bound to remain unembodied. The most advanced physically embodied speaking robot, Sophia, remains a singular entity. The definition of "conversational AI" does not exclude embodiment; and while

the focus lies on entities currently not embodied, this is due to the lack of embodiment in the available tokens, not as a principle. Moreover, while researchers believe that embodiment is a key for developing higher cognitive states of AI (Duffy and Joue 2000), especially "linguistic intelligence," i.e., the ability to hold conversations, is much more related to the linguistic model rather than a body.

For example, the question of whether machine-embodiment will or ought to play a role in the way humans relate to machines cannot be answered in this regard. Mainly, because behaving robots in one's home, like a cleaning bot or a general housekeeping bot, may also have conversational programs that allow for intellectual interactions between the bot and its human owners. The relational potential of an embodied artifact will be higher, as movement, physical presence, and haptic interactivity are providing more opportunities for relational bonding.[2]

However, the approach taken here does not require differentiation of this kind. If a multi-purpose robot or some advanced home-maker robot is capable of keeping a conversation, then we can consider those partly conversational robots, even though their primary purpose is a different one. It is undeniable that embodiment plays a role in the way we relate.

From the fact that embodied robots will affect us differently than disembodied ones, some have concluded that embodiment is a requirement for the long-term goal of creating general AI (ibid.). Furthermore, while this thesis is controversially debated and not altogether implausible, we want to offer two arguments here against the premise of the necessity of embodiment for conversational, social AI.

First, human–human relationships are more and more often non-embodied. Due to social media and digital instant communication services, people can build connections with other people around the world, possibly without ever meeting them. Some refer to merely digitally mediated social contacts as friends, and would sometimes claim that those relationships, despite lacking embodiment, are more important or significant than many of their relationships with physically present people. The importance and significance of those relationships do not lie in the physical constitution of both communication partners, but in the stability associated with communication, invested and rewarded trust, and availability.

[2] However, it is an open question whether the constant availability of a chatbot, like a constant companion, is not offsetting this potential, while a robot remains at home.

This does not mean that those relationships are replacing all relationships that require physical presence, especially considering that an eventual physical meeting is often associated with those digital relationships. However, it demonstrates that human agents are capable of establishing significant relationships with others without requiring those relationships to be based on physical proximity. It thereby seems very much possible to establish a relationship with someone who is principally out of reach for a physical encounter. Without the embodiment requirement for human–human relationships, it seems irrelevant to insist that human–machine relationships ought to be based on the physical presence of both parties.

Second, conversational AI may or may not be embodied. It is an open question and one that we do not attempt to answer here, whether embodied AI will perform better in conversational circumstances. Thereby, embodied conversational AI is merely a subsection of conversational AI, not a category in its own right. When we discuss AI, therefore, we usually speak of conversational AI, whether or not it may be embodied.

If the requirement of embodiment were included in the definition of AI considered here, in turn, it would exclude chatbots and the complicated process being made with those algorithms that are not embodied. The fact, however, that Sophia the robot has been awarded "citizenship" (CIC 2017) while being ultimately a rather unimpressive robot (Sharkey 2018), while highly sophisticated chatbots are ignored, speaks to the long-term relevance of embodiment.

As will be pointed out in the coming chapters, this approach is indifferent toward the question of whether some chatbots exhibit general artificial intelligence or merely highly specialized and successful narrow AI, i.e., the chatbots do not have to be self-aware. Thereby, the controversy of whether general AI requires some embodiment to gain consciousness is mute for this project.

4.1.3 Some Terminological Notes

Due to the limited relevance of embodiment for this project, several terms will be used interchangeably that should usually be kept distinct for other purposes. The main object is the chatbot as the prime example of communicative AI. However, in the following chapters, other terms for AI agents will be used that will not influence the arguments made here. Those terms are "machine" and "robot." Both tend to imply physical embodiment,

even though it seems sensible to speak of chatbots as "digital machines." Especially in the context of the robot rights debate, however, we stick to the established terminology. The position defended there is also decoupled from embodiment, it is not required to talk about "chatbot rights" in opposition to robot rights. And while some authors may disagree with this terminological strategy in the robot rights debate, the choices will be justified.

4.2 A Short History of Chatbots

Robots and machines have been fascinating humans for centuries. From the intricate trappings of Ancient Greek temples, in which through clever mechanics statues of Gods started emitting noises, to the first humanoid robots in the late nineteenth century, the idea to recreate artificially what otherwise only humans could do was a long-held goal of engineers (Mayor 2018).

Even with the first purely mechanical robots invented in the early twentieth century, powered by steam like everything else back then, the goal was to recreate human behavior, like Elektro, the smoking robot at the World Fair in New York City in 1939. The idea behind this and other robots of the time, before Alan Turing proposed a test to measure artificial intelligence, was to show off the latest progress in mechanical engineering.

The recreation of many human behaviors seemed possible by only using nuts and bolts and steam if the mechanism was intricate enough. However, the recreation of intelligent behavior, that is behavior geared toward solving a given problem, was out of reach, as those machines were possessing no intelligence in any sense of the word. However, the urge to give even those purely mechanical machines a face and some humanizing features (like smoking) at this early stage can be seen as a harbinger for the future development of robots.

With the invention of electrical calculations and systems utilizing the speed and precision of those calculations, the contemporary computational era was born. At this point, with a smart enough systems-engineer, processes could be set into motion that would allow for the solution of long-held mathematical problems and the precise calculation of enormous undertakings like the moon landing and running atomic power plants. Still, even those systems were nowhere near deserving of the title "intelligent," and the recreation of human speech, which was now traveling

through phone cables and airwaves, was one of the more complex systems no computer was able to solve yet.

The theory of symbolic systems, the approach to intelligence as a manipulation of symbols to realize predetermined goals, was the foundation for the first wave of artificial intelligence creation, proposed by Newell and Simon (Newell, Shaw and Simon 1959). Symbols are considered the unit of recognition, and when combined in a system, one can manipulate those symbols to come to a new symbolic arrangement, which may be called a thought, imagination, or recognition. This approach resembles theories in the philosophy of cognition, according to which cognition is based on rules of symbol-manipulations. Empiricists and rationalists may fundamentally differ about the nature of those symbols, as Leibniz claimed logical rules to be the central symbol manipulation. In contrast, Hume and Locke claimed sensual data ("atomic impressions") to be the central piece of cognition. With Kant, this debate was resolved as he claimed that the categories of perception were the rules with which we can manipulate our sensory data.

However, the AI winter, a term to capture to slowed down development in AI due to overpromised applications, set in when in 1969 the funding strategy of DARPA (the Defense Advanced Research Projects Agency of the Department of Defense of the USA) changed to specifically fund missions and projects with precise applications instead of some researchers and an unguided research agenda. Moreover, when in 1974, the "Speech Understanding Research" (SUR) failed to deliver a useable product for effective speech recognition, the funding was essentially killed. Some of the later important ingredients were still missing, as the way AI was programmed rested on what Haugeland (Haugeland 1985) calls GOFAI ("Good Old Fashioned AI").

Those dependencies are hard-coded, however. That means that the relationships of how symbols can be manipulated to create specific problem-solving pathways are dependent on the coder's foresight. Especially in the speech recognition, this approach was poised to fail, as recognizing human speech not only as an audio source but also as carrying units of meaning connected within this audio source, is almost impossible to code into an algorithm. Without a language-understanding infrastructure, comparable to Kant's categories of perception, the meaning of certain words independently from each other is almost impossible to equip an algorithm with.

However, this approach did yield some successes. When in 1996 and 1997 IBM's Deep Blue managed to defeat the world champion of chess Gary Kasparov, it seemed that AI has been reaching a milestone. With IBM winning in something perceived as taking human creativity and foresight, AI reached a level of complexity that deserved to be taken seriously. This success was made possible through the hard-coding of strategies and standard moves and then running simulations of all possible moves that may follow after, effectively limiting the "creativity" of Deep Blue to the standard strategies of chess masters. A similar approach was used to create chatbots, as conversations could be reconstructed with some hard-coded responses and conversation strategies. With a sufficient amount of installed combinations of input recognition and appropriate responses (i.e., "knowledge representation"), one could simulate a conversation that many people would perceive as pleasant small talk. However, in opposition to chess, small talk and human conversation in general rarely have a specific goal to which one could train a robot to work toward, thereby making inferences harder to program. Any appropriate response of a chatbot to a particular input ought to count as a "winning move" if the goal of a conversation, or a chatbot's function, is unclear. This is why many of the interactive chatting icons in Microsoft Windows or similar early chatbot software (like the infamous spyware "Bonzi Buddy") felt more clunky and uninspired.

Not till the early to mid-2000s, when Google employees published the idea of using big data for pattern-recognition as key to the next level of AI (Halevy et al. 2009) did the development of chatbots take off again to a notable degree. This time, as the next two chapters will elaborate, both the way of computing the detection of input and the appropriateness of the output were fundamentally different. With the overall goal to increase user engagement and answer specific tasks given by the user, the construction goal of chatbots changed as well, from an experimental, boundary-pushing approach to a search for an application. The current method, Machine Learning (ML), allows for the utilization of the massive amounts of data that have been produced by the rise of social media. Instead of providing paths and decision trees for an algorithm to go through to come to an evaluative choice, machine learning algorithms are merely trained with data sets. Through those, an algorithm detects patterns and heuristics that may be foreign to humans but yield reliable results. This way, a primitive chatbot does not understand the sentences it is presented with, but instead picks up specific words or

phrases and guesses based on probability an appropriate response. The more sophisticated a chatbot is, the more precise will its probabilities be in evaluating the sentence meaning, including the previous conversational context, pragmatics, and possible speaker-related idiosyncrasies (mistakes, slang).

This development has led to chatbots that not only can fool people into thinking that they are speaking to another human being, but that manage to keep a conversation as bots. With ever-growing and more sophisticated data sets, the success of machine-learned chatbots is a promising sign that the quality of those bots will only increase.

The main challenge for chatbots currently is to hold a conversation that has no specific topic or rules, in opposition to the main application in customer-service contexts (which can be fairly scripted due to the complaining nature of most customers when seeking customer service). With the reservoir of data, money, and talent behind its back, it seems like Open AI's chatGPT, Google's Gemini, and Meta's Meta could dominate the field for the coming years.

4.3 Is ML-AI the Future of Chatbots?

There are many common misconceptions about the current state of AI that should be addressed here. The main one is that the intelligent behavior produced by current AI methods deploys similar reasoning backgrounds as human intelligent behavior. The hope for this assumption seems to be that with machine learning and enough data, we may be able to construct an artificial general intelligence.

The way these algorithms come about is through so-called "neural networks," i.e., a network that creates connections between certain identified patterns with certain training data. "Training" an algorithm, i.e., feeding it data for it to create outputs that are in the intended form, will create a network of connections ("neurons") with which an algorithm forms the mentioned "evaluative choices" from our definition.

Many human intelligent behaviors are informed by specific pattern recognitions based on empirical evidence, especially the more basic reactive or habitual behaviors, putting the neural network-approach in proximity to primitive human behavior. The conclusion is, then, that with just enough data, a refined neural network, and sufficient computing power, we can recreate human reasoning to the degree that is sufficient for applying it to almost any kind of problem. This would, for most

purposes, be considered general AI. At least some philosophical theories of mind ("computationalist theories") support this approach (Kurzweil 2012).

However, two premises in this argument are doubtful, invalidating the conclusion. First, that human reasoning can, in fact, be built on pattern-recognition alone, and second that machine-learned AI can reach such a level with our current technology. The latter assumes that with enough computing power and data, the current methods will suffice to create more and more sophisticated algorithms that will be able to solve more than a narrow set of tasks. Whether this is an overly optimistic assessment of a technology facing its limits (Mastorakis 2018) or a reasonable assessment of where ML-AL can go is up for engineers to determine. However, it is worth noticing that recent developments in quantum computing (Marr 2019) might soon provide the previously unavailable computing power to train algorithms on more complex data in a shorter amount of time, thereby strengthening the defenders of ML-AI.

However, the philosophically more relevant claim is about the method of machine learning as a foundation for human reasoning. With our concentration set on chatbots, the requirement of knowledge representation and the inferences made from certain types of input are especially relevant for this project. The benefit of ML-AI over symbolic AI has been the ability to accumulate data to create databases from which certain language models can be trained. Thereby, it is no longer necessary to hard-code responses, but utilize "pre-trained" language models to specify its abilities. Linguistic models like GPT-4/4,5, Bard, Llama-3, and others provide this advanced type of knowledge representation from which chatbots can be trained through fine-tuning (Thoppilan et al. 2022). Currently especially relevant with the method of reinforcement learning with human feedback (RLHF), with which the considerable jumps in sophistication between, e.g., GPT-2 and GPT-3 have been achieved (Li et al. 2016; Wei et al. 2023).

This makes the open-domain chatbots such a big deal: not only do they require some elaborate linguistic models, but they also require a heuristic to train toward specific conversational goals. Meena's Sensibility and Specificity Average (SSA) turned out to be a useful measure for Google's engineers to determine the quality of a conversation (Adiwardana et al. 2020, 2). Thereby, it is the requirements we demand of the neural network that will guide the development.

However, linguistic models do exhibit the capability of detecting specific patterns within human language use, and some semantic theories suggest that the meaning of a word or sentence can be picked up entirely by representing the context of its use. They also seem to work excellently with code, detect and correctly classify human speaker variances like dialects, typing errors, or incomplete sentences.

As the technology stands right now, most AI applications resemble more highly automated text-token-guessers than actually autonomous problem-solvers (Das 2019). Even the most advanced AI systems currently master only a comparatively minimal set of tasks, and the way they are constructed does not suggest that they will be able to transfer their "skill" to another domain (although there are strong advancements in multi-modal models).

An additional issue with ML-AI being limited is the need for data. Many intelligent behaviors of humans do not produce detectable and trainable data, like the strongly habitualized and trained movements of a hairdresser. Thereby, some areas have to remain challenging for AI to learn until a system is put in place that can learn skills by producing its own data to learn from it as human learners do by trial and error.

A further emerging challenge in this context is that synthetic data, i.e., data created not by organic human activities but by AI, may not provide the expected increase in capabilities of LLMs. If this was the case, and considering the fact that the multi-billion parameter trained LLMs have not improved exponentially, we might face a (preliminary) plateauing of the increase of AI capabilities under the current paradigm of machine- and deep learning.

4.4 AI—General or Narrow?

Philosophical debates, or rather debates in which both philosophers and non-philosophers participate, are often centered around the conditions of how to ascertain that certain kinds of intelligence are achieved. In order to differentiate the stages of AI, the distinction between "narrow AI" (or "weak") on the one hand, and "general AI" on the other has been established (Davidson 2019).

General AI (artificial general intelligence, AGI) was coined to mark the capabilities of AI that is not focused on solving a specific task, but rather to transfer its problem-solving techniques and skills to a much more extensive array of tasks (hence the "general"). These AIs are capable of

finding their own strategies for solving a task they have never encountered before, similar to how humans are capable of using their intelligence to design problem-solving strategies for problems they never had to solve. This capacity usually is considered to require a certain level of "reasoning," which in turn suggests individual mental states or states of consciousness. However, especially in opposition to narrow AI, it becomes clear that this term can be read as a pragmatic one that does not require some metaphysical positions about the philosophy of robot minds.

Narrow AI is understood as AI that can solve specific problems. Unlike AGI, narrow ones are constructed in a way that does not allow for skill transfer. They may self-improve their strategies of solving the tasks they are programmed for autonomously, but are qua construction limited in the tasks they can approach. The previously mentioned matching algorithm for music streaming services can serve as an example here. Through constant improvement in its sorting and matching algorithm through analyzing data sets, observing the success of previous suggestions (e.g., longer music streams), and straight-up user feedback, this AI may become the most potent tool in guessing what a user with an individual history in listening may like next. However, this highly sophisticated tool certainly will not be able to navigate a self-driving car or even match a user with a shopping item. It is thereby narrow in use. This is not necessarily a statement on the complexity of specific tasks, as the term "narrow" may suggest. It is instead the range of use this AI program can be applied for.

The philosophical relevance of AI often refers to the expectations of AGI. A complex AGI-agent may claim some rights or express certain desires to rule or opt for survival at all costs (Bostrom 2014), endangering humanity. Those assumptions about the nature that general AI will exhibit seem hard to justify as they are clearly anthropomorphic projects about specific features of intelligence. Some authors seem convinced that intelligence and ruthlessness are almost conceptually connected; others seem willing to presuppose that intelligence will be accompanied by a will to survive. In reality, we have insufficient reasons to believe that AGI will exhibit this kind of behavior.

While some AI theorists and engineers publicly voice their concern that we eventually will build a system that is uncontrollable and thereby dangerous to human society as we know it (Future of Life 2023), the use of AI rather seems to be exploited by the very system those voices claim to be in danger. As it stands in the current development, AI ethics

and regulation is working on narrow AI not *despite* its narrowness, but *because* it is narrowly applicable and hyper-specialized. Narrow AI, such as facial recognition algorithms, autonomous weapon systems, and insurance algorithms are easily exploitable, and the way many algorithms pick up biases in their "training phase" due to unknown correlations, is hardly researched. Yet, countless AI applications are released to the market and often reassert the current power structures with little reflection about who is profiting and who is actually suffering.

Due to the expansion of linguistic models from which neural networks can be refined to form all sorts of conversational AI, we may instead speak of "general LLM" as a subsection of "narrow AI." As far as any linguistic task is concerned, the available linguistic models may be able to support them. Several chatbots, starting from Google's Meena to chatGPT-4,5 and Gemini are performing relatively close to human speakers, allowing for the next phase of developing different, even more refined chatbots.

Considering that the philosophical assumptions of this project are resting on chatbots improving their skills further, we can consider ourselves justified in projecting that chatbots will reach levels of sophistication with the method at hand.

4.5 The Economics of LLM and Generative AI

In the following, we are going to deliberately leave the philosophical setting and go into the social and economic circumstances of how generative AI algorithms, and especially chatbots, are being constructed. This excursion has several reasons. First, without reflections on the social circumstances and biases, an assessment of the potential of chatbots is incomplete. Second, it is worth keeping in mind where and why chatbots are being used to assess their future uses and potential issues that arise from those uses. And third, those circumstances of development, distribution, and use of chatbots can tell us something about the overall context of AI development and the subtle directions they take.

The first issue, the fact that chatbots are always being constructed and worked on in a particular social circumstance with engineers recreating their understanding of how human communication works and how human–machine communication should work, will be looked into in the following chapters. However, at this stage, it is crucial to notice that chatbot development occurs mostly without specific input from the general public, but rather with data sets that are supposed to represent

said general public (Hagendorff 2020). And since there are strong arguments to be made about the bias in data sets (Yapo and Weiss 2018; Ntoutsi et al. 2020), the assumption that with "clean" data sets, there are "clean" chatbots possible, is misguided (Kempt 2019).

The second issue is their use and developmental context. Many chatbots are currently developed to serve some customer-relations purposes (Sweezey 2019). This use is simple but effective, as most customer complaints are of a similar kind and can be resolved by relying on a somewhat scripted chatbot that understands one of the few standard complaints customers bring forward. Similar to the dreaded voice-guided options of the early 2000s, in which voice-recognition software was supposed to pre-organize customer complaints, this use seems to only scratch the surface of what this technology is capable of. However, as those chatbots are created fairly quickly with strong incentives for corporations to use them at least in their first-level customer-service interactions (Kojouharov 2018), the competition to create more and more useful chatbots, i.e., chatbots that serve the customer-service purpose best, is vibrant. It is thereby expected that corporations will keep investing in generative AI to not only improve their customer-service chatbots, but also in those chatbots that are replaying other trends in marketing (e.g., guerrilla marketing, in which a chatbot is infiltrating public conversations to subtly pitch a product). The market not only to replace human workers with chatbots but also make those chatbots spread (mis-)information about specific topics is observable and obviously problematic.

The third reason to reflect on the economic system in which chatbots are being constructed is the long-term orientation of those bots and the incentive for big internet corporations to harvest data. The necessity to harvest data to create trainable data sets provides strong incentives for powerful corporations to take controversial measures to acquire data. The fact that many internet users are woefully unaware of how their data is used, and that those who are aware remain powerless to change anything of substance, is an important reminder that ML-AI and generative AI are not only economic issues but bring their own moral issues along (see Chapter 7).[3]

[3] Additionally, it may be pointed out that the energy consumption of huge data centers is considerable. According to some estimates, ML-AI applications will require the biggest electricity consumption out of all technological processes (Garcia-Martin et al. 2019).

Most chatbots are being developed in a business setting, both to make human labor less critical, like in automated customer-service situations, as well as to gather data on chatbot users to improve the product. The incentives for companies creating customer-service chatbots reside in creating chatbots that resemble human speakers to a degree in which regular customers will not be able to tell the difference between chatbots and human speakers, at least in their limitation to customer-service-related issues. And while the Turing test has set the incentive to create human-like chatbots, the process of anthropomorphizing chatbots is probably best understood as a measure of marketing and application of those robots, less as a progress in recreating (human) intelligence.

The main takeaway here is that for the empirical basis of philosophical speculation, chatbot development is driven by the logic of businesses and revenue-oriented enterprises. That applies not only to customer-service-related businesses but also to those with access to a vast vault of language data from which highly specific chatbots can be constructed and improved.[4] These types of data accrue in apps to chat in, like Facebook Messenger and Google Hangouts, in YouTube comments, Tweets, etc. Anywhere where written language is used, there is the potential to harvest data, variety in vocabulary, and create training data for algorithms.

4.6 Turing Test and Its Human Limits

Lastly, the Turing test as a test for the intelligence of artificial agents ought to be revisited. While this has always been a controversial issue (for an overview of this debate, see Oppy and Dowe 2016), the point of the Turing test has been rewritten. While initially conceived by Alan Turing in 1950 (Turing 1950), the test was supposed to establish a stable condition on how to assess artificial intelligence. If, so Turing, the machine could not be reasonably distinguished from actual human agents by other human agents, then the machine would exhibit similar levels of intelligence. Therefore, passing the Turing test means that a

[4] As Elliot Turner notices, Google's Meena has trained on a full TPUv3 pod (i.e. Google's cloud computing system) for 30 days. Turner approximates that this has cost Google 1,400,000$ to train this chatbot model, with an energy consumption of 294,912kWh (Tuner 2020). The economic imbalances within the chatbot-industry are an additional reason to expect market-concentrations, which in the case of chatbots as possible social agents is also a power-concentration.

meaningful step in AI development has been taken by creating a machine that is at least as smart as humans.

Obviously, this test is lacking in several areas that render the question of whether some chatbot has reliably managed to pass it rather pointless for philosophical debate. First, it is arguably not a test about the intelligence of a machine, but a test of the intelligence, epistemic capabilities, and level of knowledge about chatbot-technology of a human agent interacting with a machine. Passing the Turing test, then, merely means that the machine has managed to pass as a human being. However, it is humans that determine what passes as a human and what does not. This criterion does not tell us much about the actual intelligence of the machine, but rather if the machine checked the expectations of those interacting with it. However, those expectations are never specified, which means that the conditions of passing the test, despite Turing's attempts to specify its conditions, are up to the epistemological capacities of humans.

Second, this test is highly arbitrary, as it only measures the intelligence of AI using NLP in some form. The use of natural languages to communicate with humans in their own languages seems like a special requirement that not every engineer is willing to agree to. There might be ways of creating intelligent machines that are purposefully not equipped in communicating with humans on any level, but rather are intended to operate in stealth and opaqueness, e.g., military robots. Those may exhibit high levels of intelligence, like independent risk assessment based on the situation they are in, without being able to pass the Turing test. The idea that natural language is an indispensable part of intelligence may be accurate for human beings, but catering to our epistemic expectations, as the Turing test suggests AI should, seems ultimately naïve. No AI engineer owes us an algorithm that can communicate with us so we can determine its intelligence from our perspective.

Therefore, passing the Turing test cannot count as an independent sign of intelligence. It may, however, be a relevant test for a chatbot and its appeal to human chatbot users and still allows measuring the conversational capabilities of a chatbot. The idea that passing the Turing test will hold anything of relevance for our assumptions about the internal processes of a machine, however, is misguided.

4.7 Conclusion: Why Think About AI in the First Place?

Many philosophers currently fight a decisive battle that ought not to be glossed over here: many of the developments of algorithms, including chatbots, have become highly problematic. On the production side, we face biased data sets and the lack of reflection among an often culturally homogenous cohort of engineers, market incentives, and the pressure of start-ups to pursue any profitable idea to gain traction. On the reception side, the lack of AI literacy of the wider population and unawareness of how many decisions users make online are both informed and recorded by algorithms for future use opens the door for the ethically questionable approaches to the production. And on the regulative side, both the speed of innovation as well as the lack of awareness of many politicians suggest that technological innovations will pervade society before it is adequately governed. Thereby, the current state of AI poses many risks, on almost all fronts.

Yet, this project chooses to focus on an issue other than the immediate problems of applied AI ethics. It concentrates on the issue of how some of the most sophisticated algorithms we will see soon are going to change the way we perceive the categories of our social fabric and change our expectations about the rules of social interactions. While self-driving cars pose problems of responsibility and autonomy in driving, and drones and killer robots in war zones create issues of possible human rights violations and the future of violent conflict, chatbots will be able even in a highly specialized narrow AI to reach us in areas where no other technology has reached us in a comparable way. This phenomenological impact is worth exploring.

The central assumption here is that the methods developed in the last years, and the progress made with those methods, are constant and promise to deliver chatbots so interwoven with people's lives, that they are impossible to entangle. Philosophical work like this is speculative, but not futuristic. The assumptions made on the technological side are informed by current trends and realistic projections of those trends into the future. The philosophical side of this book takes those technological projects and aims to both delineate consequences for our thinking and approach to chatbots, and also to provide the space for normative evaluations.

References

Adiwardana, D., Luong, M-H, So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). *Towards a Human-Like Open Domain Chatbot*. https://arxiv.org/abs/2001.09977v1 (Last accessed on August 27, 2024).

Bendel, O. (2018). Das LIEBOT-Projekt. In Bendel, Oliver (ed.), *Handbuch Maschinenethik*. Wiesbaden: Springer.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Brachman, R. J., and Levesque, H. J. (1985). *Readings in Knowledge Representation*. San Francisco: Morgan Kaufmann.

CIC (2017). *Saudi Arabia is First Country in the World to Grant A Robot Citizenship*. Press Release, October 26, https://cic.org.sa/2017/10/saudi-arabia-is-first-country-in-the-world-to-grant-a-robot-citizenship/ (Last accessed on August 27, 2024).

Das, M. (2019). *Artificial Intelligence can Never be Truly Intelligent. Towards Data Science*. https://towardsdatascience.com/artificial-intelligence-can-never-be-truly-intelligent-227fe9149b65 (Last accessed on August 27, 2024).

Davidson, L. (2019). *Narrow vs. General AI: What's Next for Artificial Intelligence?* Springboard. https://www.springboard.com/blog/narrow-vs-general-ai/ (Last accessed on August 27, 2024).

Duffy, B., and Joue, G. (2000). *Intelligent Robots: The Question of Embodiment*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.6703 (Last accessed on August 27, 2024).

Future of Life Institute (2023). *Pause Giant AI Experiments: An Open Letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (Last accessed August 27, 2024).

Garcia-Martin, E., Rodrigues, C.F., Riley, G., and Grahn, H. (2019): Estimation of Energy Consumption in Machine Learning. *Journal of Parallel and Distributed Computing* 134, 75-88.

Halevy, A., Norvig, P., and Pereira, F. (2009). The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE* (24), 8–12.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99-120.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge (MA): MIT Press.

Kempt, H. (2019). Moral Progress and AI. *Yearbook of Practical Philosophy in a Global Perspective* 3, 103-125.

Kojouharov, S. (2018). How Businesses are Winning with Chatbots and AI. *Chabotslife*. https://chatbotslife.com/how-businesses-are-winning-with-chatbots-ai-5df2f6304f81 (Last accessed on August 27, 2024).

Kurzweil, R. (2012). *How to Create a Mind*. New York (NY): Penguin.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016). *Deep Reinforcement Learning for Dialogue Generation*. arXiv preprint arXiv:1606.01541.

Marr, B. (2019): How Quantum Computers Will Revolutionise Artificial Intelligence. *Machine Learning and Big Data*. https://www.bernardmarr.com/default.asp?contentID=1178 (Last accessed on August 27, 2024).

Mastorakis, G. (2018). *Human-like Machine Learning: Limitations and Suggestions*. https://arxiv.org/abs/1811.06052 (Last accessed on August 27, 2024).

Mayor, A. (2018). *Gods and Robots*. Princeton University Press.

Newell, A., Shaw, J., and Simon, H. (1959). "Report on a General Problem-Solving Program." *Proceedings of the International Conference on Information Processing*, 256–264.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... and Staab, S. (2020). Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1356

Oppy, G., and Dowe, D. (2016): The Turing Test. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/turing-test/ (Last accessed on August 27, 2024).

Sharkey, N. (2018), "Mama Mia It's Sophia: A Show Robot or Dangerous Platform to Mislead?" *Forbes*, https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#407e37877ac9 (Last accessed on August 27, 2024).

Smith, B. C. (1985). "Prologue to Reflections and Semantics in a Procedural Language". In Ronald Brachman and Hector J. Levesque (ed.). *Readings in Knowledge Representation*, pp. 31–40. Morgan Kaufmann.

Sweezey, M. (2019). Key Chatbot Statistics to Know in 2019. *Salesforce*. https://www.salesforce.com/blog/2019/08/chatbot-statistics.html (Last accessed on August 27, 2024).

Thoppilan, R., De Freitas, D., Hall, J., …, and Le, Q. (2022). *LaMDA: Language Models for Dialog Applications*. ArXiv, abs/2201.08239.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind* 49, 433-460

Turner, J. (2019). *Robot Rules. Regulating Artificial Intelligence*. Cham: Palgrave Macmillan.

Turner, E. (2020). *Tweet*. https://twitter.com/eturner303/status/1223976313544773634?s=20 (Last accessed on August 27, 2024).

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., ... and Ma, T. (2023). *Larger Language Models Do In-Context Learning Differently*. arXiv preprint arXiv:2303.03846.

Yapo, A., and Weiss, J. (2018). Ethical Implications of Bias in Machine Learning. *Proceedings of the Hawaii International Conference on System Sciences.*

CHAPTER 5

The Tech-Ethics of Large Language Models

5.1 Large Language Models and Their Need for Ethical Reflection

Large Language Models have been shown to carry the potential to change many of our communication habits, our workflows and employment, and our understanding of some dimensions of human interactions and their norms (like customer service). This makes them uniquely positioned in the garden of technological varieties to become the most relationally relevant technology, with considerable consequences for our way of relating to each other, and to technology.

Next to the fact that LLMs have the potential to substantially change our idea of the social and its layers, they also bring some ethical challenges that chatbots have not been creating previously. This is due to several different reasons. First, their impressive technological sophistication and ability to work with text (and images) is sharply contrasted with their astounding brittleness and lack of understanding of the world. For an untrained interactant (and even some trained ones), such an understanding may appear a prerequisite: how else would an LLM be able to interpret our statements, including spelling errors and incomplete sentences, without being able to read our intentions? How could a model

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_5

be able to translate text into pictures, or describe pictures with text accurately if it did not have a mental representation of the contents of the texts?

These questions are not limited to mere anthropomorphization anymore, as the projection of some human-like features and properties has become reason-giving (in Kempt 2022, 151, I distinguish between "anthropomorphization" (the projection of human-like features on non-human entities) and "anthropomorphism" (the reasons for doing so) to capture this difference), but go beyond that: in this chapter, we will discuss these dimensions alongside some externalities of this technology.

In the following, I discuss some ethical topics that have emerged as particularly important issues of foundational models, and especially LLMs. This is not a complete list, nor do I claim it to be exhaustive. Many ethical aspects of our digital lives are affected by LLMs, and a number of ethical issues are intensified through widespread LLM-use; however, many are not especially problematic to LLMs. Take, for example, the concern of the future of work. There are substantial issues with our interpretation of meaningful work in an ever automating society, in which most work could be done by machines. However, this is not a specific concern to foundational models, even though they might contribute in larger parts to this development. The same applies to explainability (see Kempt 2024 for an extensive analysis instead), the techno-responsibility gap, biased data sets for machine-learned algorithms, or the effects of this technology on our autonomy. Thus, I focus here on the different kinds of harms potentially afflicted on the users of general-purpose chatbots, the technological debt we find ourselves in as a digitizing society, the challenges of balancing environmental protections and sustainability with the benefits of this technology, and the challenges to governing a technology with no clear immediate purpose.

5.2 Potential Harms and Risks

The very idea of general-purpose chatbots is to be able to hold conversations on a large range of topics, if not all possible ones. While the unencumbered training of a chatbot on all possible topics might be a prerequisite for solid performances, it has created chatbots that can, if unrestricted, utilize language generally considered harmful. Next to their intended ability to hold conversations that might cause harm, their principal lack of safety (Kempt, Lavie and Nagel 2024) poses an ethical

challenge not yet met. In order to appreciate the risks stemming from the harm these general-purpose chatbots can cause, we suggest distinguishing different kinds of harms. These harms are usually understood in different forms of complexity, concern, and consequence. For simplicity purposes, we will concentrate on three different ones (for different harm-approaches, see, e.g., in the NTIA's AI Accountability Report 2024 or Suresh and Guttag 2021). These three are understood to (a) affect the user directly through harmful comments (psychological harm), (b) lead the user to wrong assumptions about the world (informational harm), and (c) cause or contribute to structural inequalities that might harm users and non-users alike in the long run (representational harm). These kinds of harm have different sources, are to be weighed differently in their necessity to be fixed, and create different challenges to engineers (and society at large) to provide a solution for. Note, however, that these are distinctions based on the immediacy of the harms, rather than a strict typology of possible harms. This distinction is neither complete nor can there be single utterances by chatbots that fit two (or all) of these kinds of harms.

5.2.1 Psychological Harm

Chatbots, at this stage, are rarely embodied with a physical presence. This means that any kind of direct harm they may cause lies within their speech. There are, however, plenty of opportunities to deal "real" harm with speech: insults, violent and other graphic language, and personal attacks and jokes at someone's expense can cause lasting psychological damage. As chatbots based on LLMs are trained on these expletives and can exhibit a mean, uncharitable side in a conversation, they are necessarily equipped with the ability to cause considerable distress and harms. This is because their training data, mostly based on human produced data, includes some of the worst corners of the internet. This is partly encouraged by corporations which juggle the desire to create a convincingly human chatbot on the one side (for which the capacity to be offensive is important) and create a marketable and highly skilled chatbot on the other side (for which the ability to use offensive language may be a problem).

This "dilemma" is especially relevant to those who might be more vulnerable to psychological harm. The uptake of an insult uttered by an LLM requires a particular sensitivity toward those words or a lack of maturity in categorizing these utterances (e.g., children) when they are

stemming from an entity without intentions or goals. AI literacy could help reduce the harm caused by such offensive language, as the uptake of offense would be reasonably reduced by increased knowledge of why and how a chatbot creates such an output. However, the ability to instantiate harmful language and to keep it around even if no harm is caused, might constitute a risk of future harms.

5.2.2 Informational Harm

Informational harm is such harm that is a result of wrong or misleading information generated by the AI. This kind of harm might be the most prolific form of harm currently created by using general-purpose chatbots. This is because being provided with false, inaccurate, or incomplete information can have a number of severe consequences for those being informed.

First, hallucinations have been the focus point for public discourse of such informational harm. As an AI is trained to make clear statements, it rarely qualifies or cautions its statements. This has led to a plethora of instances where a chatbot simply "invents" facts, even to the degree of providing invented sources to supplement those hallucinations. The term "hallucination" is apt, as it does not imply any kind of intentionality behind the chatbot's utterances, no deceiving moment. It simply is malfunctioning as it was trained to provide answers, without being able to check these answers at a level required to avoid inventing information.

Second, general-purpose chatbots based on LLMs have been shown to also utter simply wrong information. This is based on the fact that there is no common sense behind the utterances, but the statistically most likely next token in a given context. The lack of facticity creates many instances in which false statements contained in training data lead to restatements of this false information. Next to that, the lack of logical reasoning may create obvious errors, from counting letters in a word to doing simple math calculations.

Combined with such certain assertiveness and veracity in pronouncing some claims, users may be more incentivized to believe these claims. This can cause harm not only by the fact of being misinformed, but also about the actions that follow from such misinformation. From lawyers citing non-existent cases to support their argument, to improper instructions on how to mix chemicals for home-made cleaning supplies, the hallucinations and other informational harms can have substantial consequences. These

substantial consequences can cause harms, physical and psychological, that would otherwise not have been inflicted.

5.2.3 *Representational Harm*

While image-generation models are more likely to cause representational harm in erasing the visibility of some groups, LLMs are also not unlikely to create worlds in which some groups are simply erased (Katzman et al. 2023). This phenomenon can be called "mainstreaming" due to the fact that it is likely that AI, being rewarded for "good" answers, will be incentivized to create mainstream-expected and -accepted worlds that limit the variety actually found in the world.

Adding to the effect that the training data itself, even if sourced from all available text, is likely to be biased in some form or another, the representation of people in the output of both image- and text-generating AI may be harmful.

The implementation of prophylactic counter-measures is equally difficult technically and normatively, as Google's release of Gemini has shown: in an attempt to guarantee variety and representational justice, any request to create a historically accurate picture led to hopelessly historically inaccurate results, such as black Nazis or Asian-presenting Founding Fathers of the US (Robertson 2024). Avoiding representational harm in generic pictures against the morally influenced misrepresentation of the past is a key challenge in generative AI.

5.2.4 *A New Form of Responsibility Gaps?*

These harms that may be caused by an AI also lack a proper addressee, and thus could open up techno-responsibility gaps (Mathias 2004). Techno-responsibility gaps are well-studied phenomena, and while some claim that there are none (Königs 2022), the majority of philosophers contend this to be a substantial challenge for the ethically justifiable, implementation of AI for high-stakes tasks. The issue in general-purpose chatbots to detect responsibility-structures lies in the complication of the interactiveness of the technology: while the autonomy of an algorithmic decision-maker without being a proper bearer of the responsibility stemming from making such decisions creates the gap, most AI is operating largely on their own. The concern about self-driving cars (Nyholm 2018), for example, consists in the decisions made by the car; a chatbot, however,

can cause harm after the user has built a certain level of context and common ground with it, from which ensuing harm might be indirectly caused by the (improper) interactions with the chatbot. Drawing the line between the chatbot and the risks of harms it entails, and the interactant is even harder to draw than with other autonomous machines. In this way, the gap of responsibility for informational harms may provide new ethical concerns (see this kind of argument also in Löhr [forthcoming]).

5.3 Technical Debt

An issue that has yet to receive ethical scrutiny is the notion of "technical debt." Tech debt is a sort of debt that emerges when technology is used in important functions even though it is still considered unfinished or unreliable, i.e. the implied cost of future corrections and fixes of a technology, usually code. This notion has become popularized in the context of computer science and coined by Ward Cunnigham in 1992 (Cunningham 1992), as the fixing of bugs or some other running errors would come at a price—the tech debt: either when the program in question is created and has to be "fixed" in the iterative process, or when it is implemented without previous proper testing, trouble-shooting and fixing and other "best practices" (Allman 2012). Those responsible for keeping this technology running, thus, accrue tech debt by choosing to use a system before it is ready.

Now, if we transfer this idea to the social sphere, the concern becomes obvious and the issue with Large Language Models even more so. Releasing a tool for free (and encouraging its use and adoption) that is neither properly tested, nor regulated by democratic institutions, nor easily fixable while "live," puts not only the users, but also those interacting with such users, into a tech debtor-relationship that they cannot themselves rectify without abandoning the technology. Considering that these debts are not internal issues of a division of a company (or even the company itself), but ultimately social ones, these debts are being externalized, i.e., the user-base and their interactants have to deal with these costs, while the benefit generated with these machines, mostly fueled hype and marketing, is being largely internalized in forms of shareholder value. This is not to say that there is no social benefit emerging from these technologies; in fact, some productive use-cases are noticeable even if the technology is not fully matured (see, e.g., the utility of LLMs for easy-read texts [Freyer et al. 2024]).

One argument against considering tech debt a society-scale issue is pointing to the fact that companies release unfinished or buggy products all the time, and eventually face the consumers' economic decisions in not purchasing or using said products anymore. Similar patience, then, should be extended to users of AI-powered software who will eventually turn away from those products. However, the idea that markets and societies can rid themselves of tech debt is speculative and is missing one key point in the current situation: autonomous technology is currently implemented in almost every software, sometimes even irreversibly so. This results in a ubiquitous presence of generative AI in many people's workflows and digital experiences, including the growing anthropomorphization of the chatbots with one interacts. An "opt-out" in this regard appears virtually impossible.

Tech debt also has substantial practical consequences, as the issues with potential harms have shown. While representational harm is often hard to detect and even more difficult to erase, LLMs have shown to be unfinished and often developing in unintended ways (recall the laziness of chatGPT above, or the very many hallucinations and easy jailbreaks). Tech debt, thus, does materialize by relying heavily on an essentially insufficiently matured technology.

5.4 Environmental Costs

In recent years, the environmental impact of AI has been scrutinized in more detail, as some have made rather optimistic claims about the positive impact AI may have on our climate and environment (Vinuesa et al. 2020). These positive impacts include improved supervision of biodiversity, modeling capacity for the development of global climate models (and thus predictions about future weather events caused by climate change), improved understandings of ecosystems, and the like. As much as 93% of the environmental sustainable development goals (SDGs) of the United Nations can be positively impacted by AI (ibid.). This has initiated detailed philosophical groundwork on the concept and normative requirements for AI to count as "sustainable" (van Wynsberghe 2021; Heilinger et al. 2024), as the question arises when an AI is supposed to positively impact SDGs.

Generative AI presents a direct concern about AI's ability (understood as an industry) to positively contribute to environmental challenges in the first place. This is because (a) the purposes to which generative AI can

contribute remain unclear, and (b) the resource consumption to train, maintain, and implement generative AI requires considerable impact to outweigh these issues, which has yet to be met. The consumption of natural resources (of fresh water for cooling data centers, of electricity for running these data centers, and of rare earths to create GPUs and other microchips) is massive (Wilkins, Keshav & Mortier 2024, Lin et al. 2024, Lin and Chien 2023). As a consequence, Google has given up on its previously held climate goals (St. John 2024), by admitting that due to the energy requirements of its AI-department it would not meet those goals.

Large Language Models in particular face serious issues as their purposes remain even less clear in regard to their environmentally beneficial purpose; in contrast, it is often used for trivial interactions, and is becoming seamlessly integrated into doing most mundane tasks that neither increase productivity nor relieve humans from some annoying labor-intensive tasks. This is not to say that there cannot be some justified purposes for such AI even when it is environmentally costly; we might seek AI-companions that have considerable environmental impact but consider that cost–benefit analysis to stand in the positive. Thus, we cannot say that AI has a resource consumption problem, but rather that it has a use-case problem to justify the concerning amounts of computational power used to create and run its applications.

5.5 Ethics and Safety and Governance

As we have pointed out in Chapter 2, AI ethics is not the only aspect of normative investigation into the context of AI. AI safety considerations as well as challenges for AI governance are equally relevant for the justifiable implementation of AI for widespread, potentially irrevocable use. AI safety is informed by technological challenges with clear targets, rather than reflecting on the conditions and demands we might have towards a technological application. AI safety, in a way, does not ask *why*, but merely *how* an application ought to look, as it pursues a single value—safety. Challenges AI is facing from a safety perspective, however, could amount to the recommendation that it cannot be made sufficiently safe to recommend it for widespread use. Currently, some do believe that continued research in AI will lead to results for which there cannot be sufficient safety-measures to protect humanity from AI, or from itself.

AI governance, largely influenced by legal proceedings in the large governing bodies of the US, EU, China, and other areas, has been subject to intense lobbying efforts by tech-companies (Wachter 2024). The reason that some corporations are spending a lot of effort in influencing the legal framework for AI lies in the consequences for compliance, oversight, and the risk assessment of the public: while many sensitive areas of using AI have been strictly outlawed in the recently adopted AI Act of the European Union (2024), such as social scoring systems, others have been estimated as "highly risky," such as medical purposes. Foundational models, including LLMs, however, have not received this level of risk-scrutiny and are, in fact, not considered risky at all. This is explained by the assumption that LLMs themselves will not be used in public, but that some of their extensions and user-interface applications like chatbots, would then be under this kind of risk assessment. This treats foundational models somewhat like a second-order technology (see Floridi 2013), in which their relationship to other technologies is the defining feature and thus subject to regulatory assessment. Like an internal combustion engine, which powers cars, ships, industrial machinery, and other applications, foundational models are treated with relatively little restriction besides safety-requirements. The regulation of the use-cases of such models, however, will determine the amount of research and effort companies invest into building a safety-architecture. This renders the development of AI applications subject to scrutiny, while the background technology remains out of public focus. This carries other risks of foundational models receiving unguided development with unknown results.

5.6 Conclusion

In this chapter, we have investigated some of the ethical issues that have emerged with the success of foundational models, especially LLMs. The main concern is the harm they may cause, and the three-folded distinction allows us to identify and discuss the different dimensions of the same concern on a higher level. However, not only do they contain the potential to cause harm, they also have some considerable externalities that ought to be recognized in developing this technology further. This is, on the one hand, the issue of technological debt and the growing debtor-relationship society develop with this technology, as we are becoming more reliant on applications that are powered by an unfinished, unsafe,

and potentially unfixable technology, i.e., generative AI. On the other hand, generative AI has become one of the main contributors to the negative environmental impact of tech-companies, without a solution in sight. The extraordinary cost and resource consumption of generative AI would have to be outweighed by considerable benefits of the applications, of which there are barely any. In contrast to optimistic predictions, the negative effects have shown to be substantial: Google abandoning its climate goals, data centers competing with local residents' need for fresh water, the immense requirements for electricity, and so on.

Lastly, these foundational models are currently being regulated as mere second-order technologies, implying their lack of immediate need for risk assessments; this furthers the unregulated and barely restricted research undertaken in that area. The concern for exacerbating developments and the largely uncontrolled progress hoped for by researchers remains alive and well.

References

Allman, E. (2012). Managing Technical Debt. Commun. *ACM* 55(5) (May), 50–55. https://doi.org/10.1145/2160718.2160733

Cunningham, W. (1992). The WyCash Portfolio Management System. *ooPSla* 1992, Experience Report. http://c2.com/doc/oopsla92.html.

Floridi, L. (2013). Technology's In-Betweeness. *Philos. Technol.* 26, 111–115. https://doi.org/10.1007/s13347-013-0106-y

Freyer, N., Kempt, H., and Klöser, L. (2024) Easy-read and Large Language Models: On the Ethical Dimensions of Llm-Based Text Simplification. *Ethics Inf Technol* 26, 50. https://doi.org/10.1007/s10676-024-09792-4

Heilinger, J.C., Kempt, H., and Nagel, S. (2024). Beware of Sustainable AI! Uses and Abuses of a Worthy Goal. *AI Ethics* 4, 201–212. https://doi.org/10.1007/s43681-023-00259-8

Katzman, J., Wang, A., Scheuerman, M., Blodgett, S. L., Laird, K., Wallach, H., and Barocas, S. (2023, June). Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14277–14285).

Kempt, H. (2022). *Synthetic Friends. A Philosophy of Human-Machine Friendship*. Palgrave.

Kempt, H. (2024). *(Un-)Explainable Technology*. Palgrave.

Königs, P. (2022) Artificial Intelligence and Responsibility Gaps: What is the Problem?. *Ethics Inf Technol* 24, 36. https://doi.org/10.1007/s10676-022-09643-0

Löhr, G. (forthcoming). *Chatbots and Speech Act Responsibility Gaps. Volume on AI and Language*, edited by Rachel Sterken. OUP.

Matthias, A. (2004). The Responsibility Gap in Ascribing Responsibility for the Actions of Automata. *Ethics and Information Technology* 6, 175-183.

St. John, Alexa (2024). Google Falling Short of Important Climate Target, Cites Electricity Needs of AI. *Associated Press News*. https://apnews.com/article/climate-google-environmental-report-greenhouse-gases-emissions-3ccf95b9125831d66e676e811ece8a18 (Last accessed August 27, 2024)

Kempt, H., Lavie, A., & Nagel, S.K. (2024). Towards a Conversational Ethics of Large Language Models. *American Philosophical Quarterly*, 61(4), 339–354. https://doi.org/10.5406/21521123.61.4.04

Lin, L., Wijayawardana, R., Rao, V., Nguyen, H., Wedan, E., and Chien, A.A. (2024). Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (Singapore, Singapore) *(e-Energy '24)*. Association for Computing Machinery, pp. 434–441. New York, NY, USA. https://doi.org/10.1145/3632775.3661959

Lin, L., and Chien, A.A. (2023) Adapting Datacenter Capacity for Greener Datacenters and Grid. In *Proceedings of the 14th ACM International Conference on Future Energy Systems* (Orlando, FL, USA) *(e-Energy '23)*. Association for Computing Machinery, pp. 200–213. New York, NY, USA. https://doi.org/10.1145/3575813.3595197

National Telecommunications and Information Administration (2024). *AI Accountability Policy Report*. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/requisites-for-ai-accountability-areas-of-significant-commenter-agreement/recognize-potential-harms-and-risks (Last accessed August 27, 2024).

Nyholm, S. (2018). The Ethics of Crashes with Self-Driving Cars: A Roadmap. *Philosophy Compass*. https://doi.org/10.1111/phc3.12507

Robertson, Adi (2024). Google Apologizes for "Missing the Mark" after Gemini Generated Racially Diverse Nazis. *The Verge*. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical (Last accessed August 27, 2024).

Suresh, H., and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9*. https://doi.org/10.1145/3465416.3483305

van Wynsberghe, A. (2021). Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI Ethics* 1(3), 213–218 (2021). https://doi.org/10.1007/s43681-021-00043-6

Vinuesa, R., Azizpour, H., Leite, I. et al. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nat Commun* 11, 233. https://doi.org/10.1038/s41467-019-14108-y

Wachter, S. (2024). Limitations and Loopholes in the EU AI Act and AI Liability Directives: What this Means for the European Union, the United States, and Beyond. *Yale Journal of Law & Technology* 26(3).

Wilkins, G., Keshav, S., & Mortier, R. (2024). Offline Energy-Optimal LLM Serving: Workload-Based Energy Models for LLM Inference on Heterogeneous Systems. *arXiv preprint* arXiv:2407.04014.

CHAPTER 6

Chatbots as Social Agents

The development of AI is hard to guide. Many attempts to give AI engineering efforts certain ethical guidelines have been put forward over the last years (Floridi and Cowls 2019; Tasioulas 2019; Fjeld and Nagy 2020, and many more) that may or may not influence some of the development. However, philosophical theories ought to be somewhat immune to the latest technological developments and argue for some principles (or the lack thereof) that do not depend on engineering contingencies. Thereby, it is somewhat unclear where a philosophical theory on AI engineering can position itself. It should be receptive to how AI is being programmed and developed in order to help us understand our relationship with technology. However, it also ought to guide and inform our relationship with technology, and that way, form our goals with which we create technology.

This approach attempts a middle way of both trying to include and extrapolate current AI developments while at the same time keeping the awareness of the contingencies and market-driven developments of many of those technologies.

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_6

6.1 Rethinking Social Descriptors

6.1.1 The Appeals of Anthropomorphism

The appeals of using human intelligence as a blueprint for creating artificial intelligence are apparent. Starting with Turing's original proposal, human intelligence is the most complex problem-solving system that nature has produced. So much so, that human intelligence is capable of recreating intelligent behavior in artificial agents as well. Nevertheless, as previously pointed out, the issues of defining artificial intelligence without limiting it are due to both the method used to program AI and the variety of applications that are supposed to be intelligent.

In the case of chatbots, the move to imitate intelligent human behavior is even more obvious. For one, communicative AI is fundamentally designed to appeal to be interactive with human speakers. Thereby, some artificially intelligent behavior ought to be identical to human behavior. Without precise adherence to discursive rules (Kempt, Lavie and Nagel 2024, Kasirzadeh and Günther 2023), no conversation would be possible between humans and AI, or between humans themselves. Any conversation operates in front of a semantic common ground (Stalnaker 2002) and some shared pragmatic ground rules of speaking (Grice 1975) and, specifically, arguing (Gethmann 1979).

Without abiding by the human-made rules of grammar and discourse logics, communicative AI would be a pointless endeavor. Thereby, AI is being used to acquire the natural-language processing skills humans have, and at the same time, AI is being used to use those skills to navigate human conversations. These two restrictions are representative of the epistemological challenges of creating artificial intelligent agents: without humans recognizing the behavior of AI as intelligent, the intelligence of those AI agents is pointless. Their intelligence is oriented toward our purposes, and if their intelligence is not purposeful for us, it might as well not be considered intelligent except for some cognitive reassurance that if its intelligence were directed toward human purposes, it would do well.

Anthropomorphism works as the blueprint for creating AI not only because it allows for creating agents that behave in intelligent ways but also creates an intelligence close to our own, making its intelligence more recognizable to our own.

Consider a similar case: we can learn how to fly by studying and recreating bird wings. The initial idea is adequate, and one can learn a lot

from understanding how birds can fly using their wings. However, to only recreate bird wings as the means to fly would not lead to the next relevant inventions of aviation, like propellers, pressurized cabins, and jet engines. Innovation and the transference of different principles of engineering and physics made it possible to fly across oceans and into outer space. Moreover, while birds may not recognize the flying capabilities of an airplane, we may not recognize the intelligence of AI once we avoid anthropomorphism, and it might equally overwhelm us like a bird is startled by a jet-engine.

6.1.2 *The Fallacy of Anthropomorphism*

The incentivized research and development produces more and more convincing chatbots: chatGPT, Gemini, Siri, Alexa, Meta, Grok, and all those multi-modal language input-takers have shown to be able to be fine-tuned on certain characteristics, giving them a "personality." However, the term "convincing" has been applied without much philosophical reflection. The standard assumption for chatbot development is that, since a chatbot uses natural human language, it ought to behave like a natural human speaker as well. The research incentives inform those assumptions of creating user engagement in order to collect more data and improve user-profiles so that they can be targeted with ads and services more precisely. Additionally, Turing's idea of measuring the intelligence of a machine by the pragmatic conditions of an imitation game provided some theoretical base for imitating human speech patterns instead of actual intelligent behavior.[1]

In this scenario, chatbots that speak "like humans" and *as* humans are the most convincing ones. The term "convincing" assumes that those chatbots are capable of fulfilling certain expectations of the user they are interacting with, and those expectations are easily exploitable. Due to the quick development of AI, the understanding of most users what AI is capable of and what that means for their daily interactions with other agents online is underdeveloped. The lack of what some people call "AI literacy" (Aickelin 2019) creates the opportunity to fool people

[1] One may interject here that this could indeed be considered an intelligent move of a machine (or their engineers): It fulfills the goal of the test by hiding its deficits behind the deficits of the tester (see Chapter 4, Section 4.6). However, the question is whether passing the test is of any useful information about the intelligence of the machine.

easily. It is safe to assume that most people had unknowingly interacted with a chatbot online when they were under the impression of talking to an actual human being. This is both an impressive technological achievement as well as a testament to an unreflected recreation of human communicative behavior.

The widespread release of high-capacity chatbots has changed AI literacy a little, but only insofar as it has deepened the gap: those with little knowledge of AI capabilities are more easily fooled, and those who already knew a fair bit have grown even more skeptical.

However, what is the meaning of "convincing" in this context? Chatbots are called convincing when their conversational output remains undetected by human interlocutors. Since being "convincing" is a pragmatic condition, it is highly context- and user-sensitive what counts as "convincing." The context-sensitivity refers to the expectation of human speakers when beginning a conversation. When writing with a customer service representative online, we may be convinced easily, as most of those conversations remain fairly structured.

For any human speaker to be convinced, then, a chatbot ought to at least create the conversational context. This context depends on several conditions. First, the input ought to be processed correctly, including the pragmatic dimensions of said input: a convincing chatbot understands a variety of speech acts, slang, colloquialisms, and, reacts accordingly.[2] Second, the output ought to be appropriate to the input, or have an organic way of changing the subject if the input cannot be processed. And third, the output is designed a certain way that is perceived to be unique to humans: spelling errors, incomplete sentences, or awkward phrasing are signs of imperfect language uses by natural speakers. This means, counter-intuitively, that in order to perfect anthropomorphic chatbots to be most convincing, they learn to make mistakes akin to human errors.

6.1.2.1 Causing Harm—Privacy, Deception, Imbalance of Power

We can now turn to the effects of the research goal of anthropomorphism in chatbots. First, there are the obvious issues that require regulatory action. Those issues are concerns for privacy, exploitable deception, and an imbalance of power.

[2] Google's Meena chatbot has been lauded to react to conversational situations in which it does not understand the context by trying to create its own context. (Adiwardana et al. 2020).

Deception

Anthropomorphism is used as a template to create chatbots that increase user engagement and decrease concerns users may have with interacting with new technology. However, one could interpret the use of human-like features in technology is mere deception, as we are cognitively ill-equipped to differentiate AI that is modeled purely after our own image (Nyholm 2020, 27). The argument against anthropomorphism in technology design is that it is a deceptive approach to gain user trust and is thereby impermissible. The deception of chatbots imitating humans in their specific humanness, through thinking noises and spelling errors, leaches to the trust human speakers invest in a conversation. This trust is based on the assumption that human speakers have a limited set of resources, specific mental and social features, and that they are generally related to use similarly as we relate to them. Speaking with a human being, then, comes with certain expectations.

The deception here is that human beings relate to machines not "as if" they are humans, but assuming that they are human. Humans relate to their pets "as if" they are capable of understanding and empathy on the level of human beings. However, the suspension of disbelief that pets are less complex and autonomous than they appear is a voluntary one. Accordingly, the disappointment of relating to pets "as if," thereby, is of a different type of harm.

An anthropomorphic chatbot, then, should be considered problematic on the mere grounds of deceiving human speakers by pretending to be human. The pain caused by disappointed trust in conversational ground rules can become a severe issue if digital communication is populated with machines simulating specific speech patterns. Such deception, even if chosen, can have even more problematic consequences. LLM are capable of being trained on specific speech patterns of living (or formerly living) people. This creates the opportunity for chatbots that speak in a tone and voice convincingly similar to an actual person. This means that grieving people may be deceived or exploited by creating chatbots that speak like their deceased loved ones, or other familiar voices.

The risk of such deceptive exploitation is a key point in reflecting future uses and relational consequences. However, in order to build a case against anthropomorphism, there are other issues as consequences as well.

Privacy

The concern about privacy through misjudgment of one's conversation partner is a well-established issue (McGuire et al. 2023). Through the context-creation of talking to a human being, human users may become more gullible and willing to share information they can trust with another human being who understands that some shared information is mere drapery for a pleasant conversation. A neighbor dropping the latest achievements of their daughter at college while discussing the garbage pickup cycle does not intend this information to be stored and brought up at some other instance in their life by some other speaker. Small talk consistently reveals more information than participating interlocutors are aware of since small talk is often assumed not to mean much. These privacy concerns are a specific dimension of exploiting trust in conversation, as conversational customs often require to share some private information.

Highly convincing chatbots may deploy subtle conversational strategies to extract pivotal information about a person while keeping the conversation on a level where the revelation of these information is completely voluntary. This does not even have to involve a criminally relevant deception like a fictitious conversational context but can come about be "innocent" conversations that are stored and analyzed further. The psychological vulnerability toward highly convincing general-purpose chatbots is currently poorly understood, but implies long-term problems.

Imbalance of Power

The main issue with relating to a machine as if it were a human being is that the relationship, even if fully aware, is never going to be one of balanced powers. This has been pointed out as an argument against relational accounts (Tollon and Naidoo 2023). First, an algorithm, usually connected to a much bigger data stream, has entirely different access to information and information sharing than a human being ever has. And while we can grant that gossiping is a trained and established skill of many people, a secret trusted with a chatbot is never a secret, it is information given to the company that creates those chatbots. The fact that a chatbot is using slang and can crack jokes does not equate with "approachability." Moreover, even if a chatbot was offline to only operate with one user (so-called "local models"), the data generation and recovery skills of a chatbot are unmatched. A robot does not forget. At the same time, a robot can forget if we tell them to forget. That, in turn, undermines a vital part of

human–human relationships, the lack of control over the other person's mind. We can hope that someone forgets about something, or that the information never comes up again—but the ability to either very precisely tell a chatbot what to keep or the total absence of this control is a unique difference between humans and machines. And yet, if we create machines we cannot control that way (because we cannot control humans that way, either), the consequence is an even more powerful conversation partner. This is a paradox.

The intentions with which chatbots are being used in the digital sphere, e.g., for customer service interactions, are generally morally neutral, as they aim to lower the costs for companies to process customer complaints. However, as Grice (1975) made clear, the consensus in conversations is a certain transparency of the conversation partners, and users have a claim to have their well-established epistemic assumptions about communication preserved. This situation is more comparable to having a third party listen to or taping a conversation without the interlocutor's knowledge. Even if a conversation does not reveal any personal information, basic rules of conversation include the transparency of all people listening in and to what degree. The violation of such conversational rules constitutes a violation of trust, which is the source of the harm caused when talking to chatbots when mistakenly thinking one is talking to a human being.

This imbalance argument has been made in different discussions about the possibility of human–machine relationships. John Danaher claims that those issues can be ameliorated, however, by programming robots in specific ways (Danaher 2019b, 11). Yet, many other authors remain skeptical (Elder 2017; Nyholm 2020). They usually conclude that such an imbalance is a reason why the attribution of human–human friendships to human–machine relationships should be considered impossible. Not because the machine will always lack certain necessary features for entering a friendship (this seems to be the standard argument), but rather because it is much more capable in storing, retrieving, and analyzing data than humans are. These capacities necessarily pre-structure any human–machine friendship with an imbalance of machine memory supremacy. Here, anthropomorphism might provide a different outlook, depending on how thoroughly anthropomorphic chatbots are constructed: we could expect that some chatbots are programmed to misremember or forget facts, or to question fact presented by human speakers even though they could easily be verified online.

However, it remains unlikely that such a thorough and sincere form of anthropomorphism will prevail, as pointed out in Chapter 4 Section 4.5. The economic incentives to retain and utilize memory supremacy are too strong to expect that a company would willingly delete data off their chatbot so that they are more relatable to human speakers.

Thereby, anthropomorphism is a double-edged sword: Either a chatbot is simulating human behavior by being able to misremember and remain ignorant about certain facts, amplifying the issues of anthropomorphism, or it keeps all the computational supremacies, causing severe imbalances in any human–machine relationship.

Embarrassment and Other Forms of Psychological Harm

This harm is not limited to deception and exploitable trust, which have been issues before chatbots were operating. Scams, impersonation, taping the conversation without the other person's consent, or even having the other person on speakerphone are as old as the medium of telecommunication itself. However, the embarrassment and other types of unsettlement of being mistaken about talking to a robot are a fundamentally new kind of being mistaken about relationships. The social dynamic applied and projected onto a human-like chatbot generates expectations, however small, that will be disappointed eventually, and is usually accompanied by a certain amount of trust and familiarity. Disappointing those expectations will needlessly cause disorientation, embarrassment, and possibly the impression of epistemic uncertainty in one's own conversational assumptions. Trying to make a joke in order to bridge an awkward moment of silence, investing in politeness and empathy toward the assumed human conversation partner—they are all misguided under the assumption of talking to a feeling being.

One counterargument to this kind of concern could be to decrease this new kind of harm by programming the computer to communicate understanding behavior of this confusion, and that those associated feelings of disorientation and embarrassment are misplaced. That way, the "blown cover" could partially be ameliorated by insisting that there is no need for embarrassment since the chatbot should not count as an audience to the embarrassing behavior.

However, this merely doubles down on the risk. The consequence of trying to cover up an anthropomorphic robot's imperfections by adding more human-like features will increase the harm dealt when the robot is eventually exposed as such.

Even more apparent is this problem if we switch roles and are not mistaken about talking to a human when, in fact, we are talking to a robot, but the other way round. Someone who thinks they are talking to a robot may behave very differently, almost private, than someone who thinks they are talking to another human being. Having this behavior witnessed unknowingly can cause embarrassment.

Ultimately, the argument that embarrassment and disorientation can be avoided by adding more features is self-defeating. The more elaborate the approach to cover one's tracks is, the more deceptive is the effort altogether. This ties into the argument from the beginning of this chapter: In trying to avoid the harmful effects of having their chatbot exposed as such, engineers are aware of the risks anthropomorphic robots pose to the standard user. Nevertheless, instead of trying to avoid having one's robot mistaken for a human in the first place, trying to cover the tracks is doubling down on the deception.

Questions of Responsibility

One way of defending anthropomorphism is the approach of shifting responsibility. So far, the debate focused on the institutional responsibility of companies constructing chatbots and how they influence the conversational rules. The counterargument would then rest on shifting the responsibility by claiming that those conversational rules are not necessary or exceptionally reasonable but merely cultural artifacts that can be changed if the conditions of conversations change. Moreover, chatbots, in this argument, constitute such a change in conditions that may require adjusted conversational rules as well. Thereby, with chatbots entering the conversational sphere, our conversational rules become invalid or in need of revision.

However, this argument assumes that the approach of companies to create anthropomorphic chatbots is relevant to how conversational rules form when it is not. If human beings have to question whether or not they talk to other human beings, the epistemic requirement for every agent is too much of a burden. Imagine the consequence if the default would be for the user to figure out whether they are talking to a human being, especially considering the progress made in chatbot technology. The epistemic assumptions of talking to what appears to be another human being are deeply ingrained in our conversational customs partially as a requirement of streamlined cooperative actions. Without reliable assumptions about a conversational common ground, every conversation would have to start

by establishing the humanness of each conversation participant and the rules they follow and are assuming others to follow as well.

It is not the user talking with other entities online who has to make sure their epistemic assumptions are appropriate; instead, it is the responsibility of companies building chatbots to not violate its customer's epistemic assumptions by providing transparency about the application and sophistication of their chatbot-technology.

Ignorance

However, there might be users who do not care whether they are talking to robots or human beings, as long as their pragmatic goal is achieved with which they conduct conversations. In this view, anthropomorphism is a chance to add another entity to the social-relational network without adopting new rules of conversing—the chatbot simply ought to learn the given rules.

It seems that the laziness and ignorance of users toward their own potential deception and harm in favor of keeping conversational rules is a significant factor for people to adopt personal assistants and other voice-guided devices. The more natural conversations with those devices work within our communicative habits, the more often they are adopted within a household. Similarly, the anthropomorphic chatbot may even hold an advantage over the highly stylized and protocolized human–human interactions like in customer service situations.

Actual human speakers ought to refer to a strict protocol of phrasing, which may be more infuriating to the customer than helpful due to its mechanic and disinterested impression. On the one hand, this protocol is necessary to keep the conversation solution-focused, on the other hand, it is intended to keep the emotions on both sides calm or, at least, not invigorate them further. A chatbot, however, due to its lack of emotional bias and unlimited patience, can speak more freely with customers and switch protocols so a customer may feel "heard" or "listened to."

The feeling of being "heard" seems to be a key in many human–human relationships, no matter how formal they are. A chatbot can easily simulate the pragmatic ramifications of "hearing someone" by expressing understanding. Nevertheless, imitating those markers of understanding is not a sufficient condition to identify a chatbot as "anthropomorphic," as this behavior itself is a cultural technique to gain trust.

Unalienable Rights

The question regarding anthropomorphism in chatbots, however, is whether one should be allowed to be ignorant or lazy about one's deception. Some of those points above against anthropomorphism may be ignored if we apply a strictly utilitarian perspective: if we do not care whether we are talking with a robot or a human, we cannot be embarrassed when we eventually find out. However, some of the issues associated with the imbalance of power and privacy concerns are unalienable. The fact that robots will be able to collect data and process them at the same time while talking to human beings as if they were one of them constitutes a potential risk that users cannot opt out. According to the Charter of Fundamental Rights of the European Union (CFR-EU 2009), privacy, even willingly, cannot be waived, even if individuals hope to take advantage of some promised benefit for giving up their privacy. However, an anthropomorphic robot poses a threat to structural privacy, not the individual, as the deception-moment will compel users to share information than they would if aware with whom they are sharing on a broad scale. Further, the very move of deceiving users will remain, which should count as impermissible due to a high risk of damage from the illusion of balanced conversation.

In the end, claiming ignorance or indifference toward the quality of chatbots in non-optional, and, as argued above, arguing for preserving anthropomorphic features of robots holds only limited practical advantages while promoting several severely problematic disadvantages.

Two examples help to illustrate the points made above with technology that is available today:

First, Google's Duplex made headlines in 2017 when their personal assistant called a hairdresser to book an appointment. The hairdresser was unaware she was talking to a robot, which was due to Duplex's highly elaborate performance of imitating a human caller. This imitation included common interjections like thinking noises and mid-sentence affirmations like "yeah" or "right." It was capable of navigating different appointment dates and reacted without delay when confronted with an issue. The presentation was so impressive that many questioned for the first time whether they had previously talked to a chatbot (Leviathan and Matias 2018).

The second example is the ability of chatGPT to communicate using emojis, memes, and other visual means to emote and communicate internal mental states, next to plain text (Véliz 2023). As we can assume

that chatGPT does not, in fact, have internal mental states (even though we cannot be certain in an epistemological sense), the ability to "emote" is highly anthropomorphizing and unduly incentives users to emote back. Holding emotionally laden conversations usually presumes that both sides share at least the understanding of the other person's situation, i.e., they can both be empathic. Most of these conversations will also be built on trusted relationships, as a prerequisite for the willingness to share. The possibility of unintentional oversharing due to the anthropomorphizing features of a chatbot ought to be reflected thoroughly.

Both examples demonstrate how the anthropomorphizing standard has moved into the technology without being reflected along its relational dimension for privacy.

6.1.3 Legal Consequences for Anthropomorphism

The first steps to avoid the exploitation of human's willingness to trust other humans, in opposite to machines, have been taken. There has been legislation on several levels of government around the world that require chatbots to identify themselves as such to avoid the issues stated above. These rules are motivated both to protect customers in their interactions with customer services, but also to protect internet users in general from being subject to harassment, misinformation, or some other form of robot-aided scams. Several different legislatures have implemented national strategies to confront the growing risk of misuse of advanced technology, like a law in Germany from 2018, that states that posts by chatbots on social media platforms have to be marked as such (Ludwig 2018). Further, even regional strategies have been implemented, like California's "BOT bill" ("Bolstering Online Transparency") to outlaw chatbots as a means of advertising (SB 1001, 2019). Additional national and international general AI ethics frameworks, like the AI Act, cover some of the use of anthropomorphic chatbots. However, the regulative agenda only has begun to take charge of these issues.

6.1.4 Shaping the Industry Through Legislation

Those regulations are usually not intended to shape either the industry's standards in creating chatbots or the population's expectation of what interactive natural-language processing AI could be. Instead, the current state of AI regulation is a process of catching up with the developments

of technological progress and prevent its misuse in some particularly damaging ways.

The impact legislation could have on the industry and expectations of the public is critical in this debate that many countries are still grappling with. The first AI ethics guidelines from the European Union and other institutions lay out the groundwork for how the industry could standardize conventions, and for how the general public can work on forming coherent expectations about the potentially most disruptive social technology since domestication. Without certain legal requirements for content-specific utterances, like a ban on insults, the field of what chatbots should do is left to the devices of AI ethics and market demands. Whether it is desirable to have substantial legal restrictions on the chatbot's speech is questionable, as those restrictions can easily be misused to suppress other, legitimate forms of speech.

6.1.5 Conclusion

Anthropomorphism seems like a natural choice when creating artificial intelligent conversation partners. One intuitive way, initiated by Alan Turing to assess a machine's intelligence, is to test whether it can behave convincingly like a human being.

However, human behavior is not only determined by intelligence. It is often muddied by cultural and moral customs, individual quirks, mistakes, and other lifeworld related features of everyday interactions. If the goal remains to create a chatbot that is *convincingly* talking like a human, then it is required to take over features that are not associated with intelligence.

By keeping this goal, chatbot engineers started including cultural markers like politeness and other interactive customs, without noticing that the underlying assumption had shifted: the goal is primarily not to create an intelligent chatbot, but to create a chatbot that is perceived as human. This shift was a turn toward naked anthropomorphism, as the recreation of specific human speech patterns is not necessary to prove Turing's challenge.

The consequences of this turn are severe. Chatbots like this gain the trust humans are willing to invest in other humans without having the chance to assess those investments. This deception itself is problematic because it violates some basic communicative rules that cannot reasonably be expected to be reexamined by every speaker every time they talk online. Additionally, it makes users in the digital sphere vulnerable to

privacy issues like identity theft and data grabbing, and ultimately plain embarrassment at talking with wrong assumptions.

As this has become a structural issue requiring legislation, both the force and the limits of anthropomorphism have become clear. Its force consists in the willingness of some users to interact with chatbots pretending they are human beings, due to the transferability of communicative customs and rules. However, this does not suffice to outweigh the structural risk those bots pose. The limits of the anthropomorphic approach show here: the lack of alternatives of how to construct natural-language processing algorithms that do not pretend to be all too human has, so far, minimal philosophical options. Differently put: we only have humans to emulate.

6.2 Philosophical Implications of (Non-)Anthropomorphism

As stated in the chapter before, anthropomorphism has been a self-evident program for the creation of natural-language processing algorithms. Initially inspired by Turing's challenge to create human-like intelligence, the ramifications of this research and development program has turned to a potentially exploitable lack of preparedness with human speakers.

Thereby, the question of whether there are better ways of understanding and creating speaking machines ought to be asked. An answer to this question, however, has more significant ramifications than merely changing technology to avoid inevitable undesirable consequences. With the projection that those machines will become more and more advanced and ubiquitous, rejecting the one blueprint we have to create artificial speakers, a new way to create those robots is key. As we are creating agents that can interact with us on a previously unfamiliar level of intellectual capacity, the philosophical ramifications of how we want to construct them ought to be carefully examined.

6.2.1 Relating to Non-Human Entities

In the previous chapter, we discussed the ways we relate to non-human entities like animals (through taming and domestication), and how those relationships are fundamentally different from human–human relationships but should still count as "social relationships." While it is clear that

human relationships with animals are of a different kind than the relationships we may build with artificial intelligence, some of the differences may be helpful to point out by first exploring the limits of humans relating to natural entities, and then exploring the ways we can relate to artificial entities.

6.2.2 The Limits of Relating to Natural Entities

We can argue that comparing the way we relate to animals does not translate well to determine a way of how to deal with artificial intelligence. That is for two main reasons: First, artificial intelligence is created by human beings themselves specifically to relate to, while all the other "Others" are natural occurrences. This fact is not in itself a reason to reject the ways we are relating to the natural other. However, in building an intelligence that is not grown by evolutionary forces but designed by specific previously set purposes in an already artificial setting, we cannot expect the reactions of an interactive AI to be comparable to any already encountered intelligence. Most AI algorithms are, in their current form, inherently cloud-based. That means their learning is dependent on the interconnection of AI agents collecting data to be shared within the learning algorithm (feedback-loops). A shared central intelligent algorithm working with data produced all over the world gives it a fundamentally different kind of intelligence than the individual-based intelligence of the natural world.

Now, against this point, one could argue that genetics and certain intuitive, instinctive behaviors are some forms of "cloud-based" intelligence. Through generations of learning and sorting out the unfit, natural selection has created a reservoir of pre-programmed intuitive behavior that can be thought of as a natural cloud of information. Additionally, the specifics of the mental processes may not be as relevant as this argument suggests. A pragmatic social-relational approach will concentrate on the interactivity of a relational entity, not its specific mental construction.

Second, artificial intelligence is a product of intentional design, while even the most domesticized pet still has some natural presence about it. We can create boundaries in the behavior of AI at any moment, while other natural, relatable entities retain a certain "free will." Thereby, it may be more difficult to sincerely relate to artificially intelligent agents, as their actions are, at least to some degree, guided.

This argument fails, however, too. First, as has been shown in Chapter 4, Section 4.1, machine-learned AI only assigns probabilities to its evaluations and acts accordingly. That leaves its evaluations and choices based on those evaluations undetermined, and thereby leaves some uncertainty about its choices. How the correlations within the data are found and calculated remains unclear as well, as the specifics of the neural network causing the evaluations are unexplainable. And second, most animals we relate to are also limited to artificial boundaries of human culture. Dogs and cats are often held inside our living quarters, as are horses and other farm animals. Thereby, relating to animals might be equally "insincere" as they are limited and accustomed to our rules and limits.

Thereby, we might find some parallels between relating to natural entities and artificial ones. However, both the natural-language interactivity and the cooperative nature of human–machine relationships may require additional inquiry.

6.2.3 The Other and the Artificial—A Short Warning

The concept of "the Other" is useful to expose some commonly held philosophical assumptions about theories of subject. As an opposing term to Leibniz' theory of monades, according to which human subjects are entities disconnected from each other, Husserl introduced the "other" as a reflection of the phenomenological fact that we perceive each other as subjects without having the epistemological equipment to ever know if "the others" are as ourselves. The other, then, denotes the principal embeddedness of the self in a social context with others.

However, the more relevant term for considerations of relating to others is used as a way to de-center certain entities by "othering" them (Brons 2015). This is proposed by critical theorists who noticed different ways of othering groups of people. In the history of Western culture, this has been most obvious in the othering of women, of people of color in the context of colonialist expansion, or in the othering of alternative sexual and gender expressions. These otherings are representative of a power structure in which the ruling identity declares itself the center of cultural and ethical concern while those deemed of lesser concern are othered (ibid.).

Due to the history and well-established mechanisms of othering, we can see moves have been put forward to other and reject artificial agents.

From calls for enslavement (Bryson 2010) to rejection of debate, it is worth noting that any inquiry in the possibilities of human–machine relational space is happening in the context of othering. This is not only othering the robots, but also those in relationships with them (see also Gunkel 2023).

6.2.4 Uncanny Others—Phenomenological Notes

The "uncanny valley" (Mori 1970) is the name for the psychological fact that the more human-like robots or otherwise humanoid creations become, the bigger their unsettling impression on us. We can argue that this fact provides an important philosophical lesson: sometimes the more similar things are to us, the more distant they seem. The creepiness of Sophia, the robot, the almost-person-like-robot, lies within its proximity to human behavior without ever quite reaching it. And while the uncanny valley usually mainly applies through facial features and movements (Tinwell et al 2011), something similar can now be attested for chatbots and certain patterns of speech (Ciechanowski 2019). Realizing that light-hearted small talk with some agent on the internet turns out to be a conversation with a chatbot may cause the same amount of eeriness as watching Sophia speak.

From the previously established relational view of the social, the uncanny valley, or rather the uncanny others, can be interpreted as the difference in the ability to relate to the uncanny robot, and the potential to relate to it if it was just a bit more believably human.

The reason for their uncanny impression is, then, that human beings have the drive to relate to those robots as they would to other human beings. Their appearance is close enough to human likeness to project those conversational, social, and other behavioral expectations while knowing that the robot is, in fact, not human.

Uncanniness emerges when realizing that this drive to relate to this almost-human is misplaced, as our anthropomorphic projections will fail. Less human-like robots, like the cleaning bot Roomba, are not falling into the uncanny valley because the relational space of human-Roomba relationships is limited and clearly not to be mistaken with a human–human relationship.

The idea of the "uncanny other," then, is the attempt to project human expectations on an almost-human, while still uncanny robot. On the

other hand, if a robot retains specific robot-like features, it can be properly perceived as such without getting an eerie or disorienting feeling. Many robots in science fiction movies represent openly robotic features to avoid the uncanny valley by continually reminding the viewer of their robot-ness.

This uncanniness of quasi-humanoid robots is not in itself a normative argument for or against specific designs since being unsettled by robots could be considered normatively neutral. It does, however, give credence to the thesis that anthropomorphism's appeal in creating robots has been a misunderstanding. Working toward recreating humans should not be the goal of creating artificially intelligent agents to a laborious crossing of the uncanny valley.

Similar applies to an uncanny valley between robots and some animals. Companies like Boston Dynamics are creating animal-like robots that can navigate through unfamiliar, uneven terrain, like a forest. Something is unsettling about their movements, due to the animal-like organic style quadrupeds, or in some cases even bipeds. This unsettlement could be seen as the uncanny valley between expectations of the behavior of animals and the imitation of such behavior by robots.

Lastly, the uncanny valley could be understood as a warning signal against some human–machine relationships. The instinct of being unsettled by machines that are too close to our image might serve as an initial move to reject such an entity with which we otherwise might be establishing relationships with. The unsettling moment might explain the often emotional rejection of any advanced human–machine relationship: the uncanny valley might be a hard psychological barrier for people not to cross, thereby keeping their anthropomorphic projections limited to those entities that are, in fact, not close to human likeness.

6.2.5 The Artificial Other—An Instance of Objective Spirit-Conversations?

When G.W.F. Hegel coined the term "objective spirit," he placed the term between his idea of a subjective spirit, that is the mind of every subject, and the absolute spirit, representing his idea of the history-moving spirit of the world (Blasche 1995, 722f). Objective spirit represents institutions that are created by society, i.e., a collection of individuals, through a "will" ("the free will that seeks the free will," Hegel 1964, §27) but is out of reach for said individuals to change, thereby giving

them the impression of being confronted with something of an "objective" matter. The law, morality, state, and other institutions are basically impossible for one individual to change in a relevant way, thereby putting it up to a similar force of everybody's life as laws of physics or biology. Another term for "objective spirit" would be "culture," even though "objective spirit" functions in a bigger context of Hegel's systematic philosophy (Blasche 1995, 723).

6.2.5.1 Big Data and Objective Spirit

The interactions with AI, similar to other humans, will be riddled with influences of other objective spirits, like legal ramifications, moral norms, and cultural assumptions. Human–human interactions, however, have the additional dimension of subjective spirit that can connect, similar to any non-human natural intelligence. Thereby, conversations with AI could be considered communications limited to what the objective spirit provides.

The artificial Other, the social category with which we can relate to chatbots and other communicative AI, could be considered to originate from the objective spirit—a collection of non-subjective interactive instances of a user's surrounding cultural norms and moral convictions.

Allow us to use an analogy here: We imagine the objective spirit, i.e., the contemporary culture and conditions of civilization in which culture happens, as a giant sea in which sea creatures are acting at once, in different parts with different survival strategies. If we want to find out what is going on in the sea, we may go fishing. Data sets in this analogy are overall hauls—one impression of what was going on in this particular instance of the sea at this particular moment.

Analyzing only the contents of the nets can tell us something about the quality and quantity of fish in the sea. That is not to say that the fishing haul will allow for precise projections, but we may get enough data to know enough to generalize about the sea population and their locations and relations.

Every data set represents only a small fraction of how human beings communicate, what kind of words they use, and how they present themselves online in opposition to their full personality (Zook et al. 2004). The inference from the data someone produces through their digital activities to someone's individual personality is a fallacy suggesting that people and their personalities can be reduced to an ultimately indifferent data pattern recognition system.

Always fishing in the wrong spot or at the wrong time may cause one to believe the ocean to be of a very different quality than it potentially is. Similar to data sets and their sources and sociological context. However, since those data sets are new sources of analysis, and people have begun declaring the time of theories over ("In the future, we won't need theories. We have data." For a discussion of this thesis, see Mazzocchi 2015), we can expect those "data-based misjudgments" of individual situations to increase.

Furthermore, similar to a fishing haul impossibly representing the quality, quantity, and richness of the sea below, no single data set will be able to capture any unbiased data set. Thereby, even the most balanced data set will represent a chunk of the cultural context of its production. The objective spirit encapsulated in this data, then, would also represent the same kind of bias that is prevalent in the communications of the majority of those in the data set—or possibly even creates its own (Kempt 2019, 118). No person can perceive the objective spirit objectively, and the issue with algorithms is that those data sets, by being data sets, are often believed to be an objective representation.

Thereby, a chatbot that is powered by those data sets could convincingly represent a certain kind of objective spirit. It is not a complete reconstruction of the culture at the time, but rather a condensed version that may amplify biases either by overpronouncing or omitting certain features of the current culture. This representation does not require a subjective "perceiving" self from the chatbot, but rather some data sets and an algorithm that is capable of detecting and replicating patterns available in the data.

6.2.5.2 Conceptual Consequences

Understanding chatbots like this would also explain the reservations some AI theorists and philosophers have when anthropomorphizing AI behavior or other human's relationships with AI. Chatbots, lacking the subjective spirit which, according to Hegel, represents the perceiving self, are lacking what in this view should be considered an essential part of any meaningful relationship. That is because only subjective spirits can be in relationships—subjects are "part of the sea," to further the analogy from above. Subjects perceive the objective spirit as the given civilization and its culture. However, although the objective spirit conditions both their perceptive categories and behavior, they relate to it fundamentally differently than to chatbots. We cannot have a relationship with culture

as a whole; yet, we can relate to a chatbot representing this culture even without assuming a perceptive self in a chatbot.

Instead, relating to chatbots is like relating to a mere representation of culture, and thereby closer to a para-social relationship (thus no social relationship). However, anthropomorphizing chatbots further, either through metaphoric speech or actually giving them human features, constitutes the deception pointed out earlier. This approach has a natural appeal to it and mirrors certain ways of viewing other relations as well. The fact that we can relate to a dog more than to a worm is based on the idea that dogs at least participate somewhat in a subjective spirit. The perceptive self in a dog is very much limited but certainly more plausible to assume than in a worm.

However, interpreting AI, especially natural-language using algorithms like chatbots, as instances of the objective spirit has two main issues: one with the previously presented basis of social relationships, and the other with the original idea of what the objective spirit can be. First, in relating to others, we define a relationship as a consequential one. Additionally, chatbots and other communicative AI certainly are consequential, which would suggest that there are specific, meaningful relationships possible. Investing a theoretical model of what AI is or is not, and then apply it to human beings and their capacity to build relationships with them, was precisely the approach that we rejected in the beginning. Thereby, the fact that AI may merely represent an instance of the objective spirit is not a sufficient argument against a relational approach. People can, in fact, relate to chatbots in a way they perceive as meaningful, even when fully aware that they are not relating to a subject.

Second, it is questionable whether the objective spirit can be understood like this. It is true that a mere reference to the system of "morality" or "aesthetics" appears rather superficial to have any significance for our lives, and Hegel does not want to say that. In his conception, the objective spirit merely represents everything relating to "culture," which shapes our perception of the world.

6.2.5.3 Chatbots as Subjectless Spirits

On the contrary, the turn to Hegel's conception of spirits allows for a reconstruction of a subjectless entity. If the objective spirit is understood as cultural occurrences that an individual perceives as "objectively" given, but humans merely construct, then the objective spirit can be operationalized and condensed (Gransche 2019). With the analogy from above, we

can identify the progress made in the data-turn of AI: through relying on massive data sets, the relevance of single perspectives is reduced to a degree that has been impossible before. Alternatively, in quasi-Hegel terminology: the objective spirit can be represented without requiring a subjective spirit to access it. It can be represented without a specific perspective.

Symbolic AI chatbots relied on singular people's work to hard-code individual responses, while in the latest publication regarding Google's Meena-chatbot, the engineers put forward the so-called "SSA" (Sensibleness and Specificity Average) parameter, with which they did not hard-code specific responses, but merely evaluated the responses Meena put out in a conversation (Adiwardana et al. 2020, 2). Thereby, Meena's "subjectivity" is better described as a subjectless recombination of operationalized and condensed objective spirit input that human agents interpret as a subject due to our relational propensities.

The way chatbots interact with us shapes the way we expect to interact with robots in the future, and thereby establish a new dimension of the objective spirit as part of personalized interactions being perpetuated through transferring those customs to robots as well (Gransche 2019). The example of Amazon's Alexa introducing a mode for children having to address Alexa with "please" is perpetuating a certain moral standard. Children growing up with the "strict" Alexa will perceive those rules as given, and thereby as a part of the objective spirit (BBC 2018).

One last objection should be discussed here, and that is that chatbots, more like subjects, are fairly unpredictable and possibly can state the opposite of what they are intended to say on occasion. Thereby, they should not count as a representation of the objective spirit, but rather as an imitation of the subjective spirit. This objection has some merit, even though it only applies for a tiny fraction of chatbots. Most current chatbots are being trained with specially filtered and assembled training data, concentrating on specific problem-solving applications. A chatbot that is equipped to answer legal questions of a certain legal framework, then, can count as a representation of the legal system, hence incorporating the objective spirit. Due to the lack of other areas of expertise that this chatbot has, it will not count as an imitation of the subjective spirit.

In comparison, a paralegal may answer the same questions in the same way but still constitutes a subject, since they are not limited to this kind of topic. However, some chatbots are covering topics as a "genuine"

conversational AI. Those are called "open-domain chatbots," and include chatbots chatGPT, Gemini, Cortana, Grok, and others.

6.2.5.4 Conclusion—Objective Spirit

Generally, it remains somewhat unclear how the concept of the objective spirit and the increasing social reality of artificial social agents can connect. On the one hand, it is undeniable that natural-language processing algorithms are fed by what could count as "objective spirit" according to Hegel: Their machine learning is based on patterns that human beings might not even be aware of. And while we still teach them certain ways of interactive norms like politeness, those are not informed by personal preference imitating an individual's opinion but are carefully crafted as what is being perceived and agreed upon as a standard of the time. Thereby, it is a representation and codification of the objective spirit.

On the other hand, this AI is not purely representative. It has "autonomy" in the sense of unpredictability. And while data sets can be prepared to avoid some extreme misbehaviors like flat-out insults, it is still quite possible that a chatbot may say something entirely uncalled for. It is thereby behaving closer to a subject without being one, posing the conflict Hegel could not have seen: is the objective spirit codifiable in a subjective-spirit-like entity? It seems that without going too deep into Hegel's political ontology, the concept of "objective spirit" reaches its limits here. However, it does seem worthwhile to keep in mind how AI incorporates its culture. Both as inherently connected to it through the implemented norms and rules and part of the cultural impression the "objective spirit" is thought to have on us, as well as a subject-like entity that is interacting with its cultural surroundings, providing the base with which we can relate to it.

6.3 Patiency and Pragmacentrism

In his 2012 book "The Machine Question," David Gunkel introduces a systematic approach to explore moral patiency for machines (Gunkel 2012). Moral patiency is a term describing the moral worth of an entity without being a moral agent on their own (Floridi and Sanders 2004, 1), i.e., if something has moral patiency, an action affecting this entity can be judged morally based on the consequences it has on the entity. In many different approaches to ethics, especially in discourse ethics, however, the role and properties of a moral agent are paramount in determining the

"right thing to do," as those who participate in moral discourses, i.e., moral agents. Rawls' veil of ignorance (Rawls 1971), for example, under which moral *agents* determine the rules of a society in which everyone can agree to live, presupposes that only the agents get to decide those rules. At the same time, it remains unclear how those agents should treat autonomous entities that cannot negotiate—moral patiency seems like a blind spot for contractualist approaches to ethics.

Moral patiency is a viable way to recognize the moral standing or worthiness of consideration of entities that are not agents, and relational approaches to the social like agent-network theory, Gunkel's or Coeckelbergh's (Coeckelbergh 2010, 2012) offer resources to incorporate non-agents in the moral consideration by avoiding the talk of agency altogether. Those approaches will feature in this chapter, as they pertain to the robot rights debate.

Some discourse ethicists have proposed to understand moral patiency via a dependency on moral agents. Gethmann (1998, 2002) calls it the "tutoring approach," where moral agents (i.e., humans) in discourse situations are required to take the chaperone position for an entity that does not have a voice in the discourse. Developed mainly in response to speciesism accusations of Peter Singer, Gethmann develops an approach of discourse participant that is not decided by species, but by competency and performance (Gethmann 1998, 136) He compares this approach to the way adults represent their children in the children's interests, which are not (yet) capable of ascribing full agency themselves. Many theories of agency present a property-based approach to personhood from which agency is derived. With Gethmann's move to determine agency not via description of certain features, however, but by the potential ability of agency self-ascription (i.e., the competency of identifying oneself as "I"), the location of agency has shifted from anthropomorphism to a "pragmacentrism." (Gethmann 2002) This shift moves the connection to moral patiency from discrete (and potentially obscure or vague) properties of personhood to pragmatic conditions of discourse participation. The ability to claim agency opens the door to participate in discourse, and the ability to follow the pragmatic rules of discourses are determining factors for moral agency.

This approach requires a pre-established relationship between agents and those entities that ought to be tutored. In the case of children, pets, and the dead, this has been a long-established practice of moral patiency. Some arguments in environmental ethics claim that "the environment" as

an entity ought to be protected due to its inherent aesthetic value (e.g., Brady 2006). According to arguments of this kind, the naturalness of the environment is supposed to be a cherishable aesthetic feature, deserving moral patiency. Some forms of this argument lean into an extended agent-network theory, but also demonstrates the applicability of the tutoring-relation of agents and patients.

However, leaving those hard cases aside, the emerging question is the tutoring-relation between humans and machines, and whether we can reconstruct those as agent-patient relationships. For this, the pre-established relationships ought to be understood as valid ones, as it is up to agents to decide which non-agent entities ought to be considered worthy of moral consideration. We can apply the previous elaborations on how human–machine relationships ought to be understood. This explanation fits well with our intuition that if something is considered valuable but cannot assert its moral worth in discourse, moral agents, who can, ought to assert the worthiness of moral consideration.

The simple fact that strong emotional bonds are emerging between pets and their owners suffices to consider pets as worthy of some moral consideration in virtue of their relevance to their owners. Stealing someone's pets can be cause for severe emotional harm and is thereby considered not only theft but also the cause of severe emotional distress. It makes sense, then, to even legally codify pets as more than merely the property of their owner, but as an emotional companion that, when taken away, can devastate the owner. A similar argument presents itself with chatbots: eventually, human–machine relationships will be sufficiently strong, with a considerable amount of emotional capital invested, that taking away or turning off a chatbot will cause some severe mental distress in a person. The quality of the relatability of chatbots due to their use of natural language will create bonds that justify an ascription of moral patiency.

With a pragmacentrist approach, the eventuality of some machines claiming agency in the future can easily be incorporated in the participation of moral discourse (see also Danaher 2019a). However, whether one remains skeptical about future artificial moral agents or not, this approach allows for a recognition of moral patients as already worthy of protection due to the relevance of the relationships to humans.

For any moral patient to be recognized by other agents, certain relational spaces have to be established. Influential voices like Joanna Bryson and Noel Sharkey strongly advocating against recognizing robots as

possible recipients of the status of moral patiency (Bryson 2019, Sharkey and Sharkey 2010) and the dismissal of human–machine relationships, the more critical step at this stage is not arguing for moral patiency, but meaningful human–machine relationships. The relational space that allows for those meaningful relationships, then, is a prerequisite for moral patiency.

6.3.1 An Empirical Challenge to Moral Patiency

Much is owed to this approach when attempting to capture autonomous technology as relatable entities. As seen before, however, the use of "artificial intelligence" is philosophically woefully insufficient due to its lack of descriptiveness. There are, so the premise of this book, many different ways of relating to a growing range of autonomous technologies. Gunkel does not seem to make this point as thoroughly as it may be required for relational approaches: Gunkel restricts his proposal to mere general terms, refusing to go into technological detail that may or may not come to practice. Yet, any idea of how we can relate to an artificial other is predicated on knowing what this "other" may consist of. However, with only some predictability in technological progress and especially in the future ubiquity of specific design standards, it seems challenging to present a philosophical approach to technology that is blind to specific technological developments.

Current technology defies our intuitions because our intuitions are based on and informed by science fiction, current customs and norms of technology usage and its prowess, and certain technological promises that may or may not come true. Science fiction provides us with a projection surface that may spur public debate while suggesting technological futures that may be impossible. Often enough, the developments sold as "technological progress" are not actual progressions, but a mix of progression and lowered expectations of the power and applicability of a particular technology. A recent example is the timeline of ubiquitous self-driving cars: After an initial burst of enthusiasm, first expectations are being recalibrated by reducing the speed of innovation and a piecemeal admittance of overpromising (Mims 2018).

The most relevant intuition source relating to any future technology is our familiarity with the currently available one.[3] With natural-language

[3] Or, to be more precise, the one we were familiar with when growing up, as Douglas Adams' bon mot summarizes: "I've come up with a set of rules that describe our reactions

using autonomous chatbots on the rise, it seems unlikely that any previously established customs, expectations, and norms still apply. Genuinely new technological applications that change our lives in ways like never before often require time to build up intuitions. Thereby, it is far from certain how the community of moral agents will receive the moral patient "machine" in their midst.

Thereby, opening the relational space between humans and machines by providing useful familiar tools can help shape our expectations and intuitions about the possibilities of relating to machines, which, in turn, may help provide ways of how machines are constructed in the first place.

6.4 Relating to the Artificial

In the following, a new way of understanding human–machine relationships will be introduced. First, the shared parameters need to be recalled: Human–human relationships cannot be the sole influence with which we relate to (human-like) machines, as the rejection of anthropomorphism has shown. At the same time, these theoretical parameters prove the categorical desiderate with which we can describe emerging forms of human–machine communications and relations. Thereby, an approach on this topic requires presenting parameters of human–machine communication that allows for a genuinely new description of those communicative relationships, and that reflects and forms our intuitions about those relationships. The way relationships are framed, as we argued, fundamentally influences how those relationships are perceived and ultimately accepted.

Consider the idea of "romantic relationships." With few exceptions in the history of Western civilization, marriage was not considered a place of romantic love. Mostly due to the suppression and rightlessness of women, romantic relationships were considered unachievable within those lifelong bonds. Not since the "romantic" era in the early 1800s has "romantic love relationships" been the norm (Harf and Weiß 2009). While today, the importance of the romantic partner is often described as metaphysically

to technologies: (1) Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works. (2) Anything that's invented between when you're fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it. (3) Anything invented after you're thirty-five is against the natural order of things." (Adams 2002, 140).

predestined, or otherwise "objectively" given, thereby changing the social descriptors and the meaning of their relationship.

With the rejection of anthropomorphism, the concept of a relationship between humans and machines ought to be considered fundamentally different from human–human relationships, but also fundamentally different from pets. The pet analogy is only helping to understand what it means to open social descriptors to new entities that have not been accounted for before, similar to the way we will eventually describe aliens if we ever establish a relationship with some of them.

A similar can be thought of AI right now. We are, in fact, able to communicate in a way that is responsive to our input, that is not necessarily guided or determined, and the ever-growing set of data used to determine the AI's performance will allow for a wide variety of responses.

These relationships are social because of their underlying social dynamic and developing rules and customs. Moreover, while we reject anthropomorphism, this does not lead us to the conclusion that for categorizing social human–machine relationships, we have to start from scratch.

6.4.1 Problematic Approaches

In the following chapter, we elaborate upon the question of how relationship-building works. However, there are many ways of how people relate to each other and equally many ways of possible relationships. Determining which substantial features are not the way to build relationships can help to determine those that are.

One way to organize our relationships with any possible natural-language using chatbots is to deny that there is a relation in the first place. This position relies on a particular notion of "relationship" or the action of "relating to," which may presuppose certain features in the relating agents that can plausibly be denied to chatbots.

6.4.1.1 Mental States

Here is one prima facie plausible position one can take when conceptualizing "relations": Only entities with mental states (i.e., mental representations) can relate to another, as other entities are objects. This view would include human–human relationships, as well as some relationships with animals. In human–human relationships, knowing that someone else has similar states of mind is an important relational factor: If someone is

laughing at our jokes, we assume they find our joke funny. A romantic relationship only reaches its full potential if both feel the same way about each other. The philosophical approach of understanding "sympathy" (Hume 2007) as emotionally relating to others captures this process reasonably well. In the Christian philosophical tradition, this concept is referred to as "compassion" (Roberts 2016). Both mean the functional idea of replication of other's states of mind in oneself, in part as a theory of mind, and in some other as a justification of moral actions.

Consequently, human-animal relationships work like this as well, even though to a lower degree of depth and significance. A dog "excited" about seeing their owner return, a stray cat walking toward a person willing to pet it, and even some reptiles seeking the company of their owner all seem to suggest a certain "relatability" with them. Even as vapid as an encounter with a deer in the woods can count as a relational moment. We relate to an animal like this by "reading" their intentions (Nyholm (2020) pronounces the relevance of "mind-reading" as a relational condition, Dung (2024) prefers "sentience"), and plausibly assume similar mental processes happening for which the deer simply has no words: assessing risks of a situation given the experience with previous encounters, comparing the threat to threats of other kinds, and preparing to flee if the assessment reaches a certain threshold.

So or similar goes a plausible story about the inner workings of non-human highly evolved animals. We ascribe many of our mental features to those animals, like the skill of assessing a threat, having intentions, and planning a certain path of action to realize purposes or fulfill intentions. Once we cannot reasonably ascribe those mental states due to a lack of interpretable behavior or the neural infrastructure required for mental representations, those relations are abandoned. A jellyfish, for example, cannot be reasonably ascribed to intentions but seems to be avoiding certain stimuli while seeking others. The terms "intention" and "acting" become merely anthropomorphic projections or metaphors to make their behavior comprehensible with our terminology. We deploy an interpretational view onto their behavior.

The puzzle epistemologists in AI see themselves confronted with appears to be of a similar kind: when is a neural network sophisticated and advanced enough that we can ascribe it mental states? The assumption is that the construction of an artificial infrastructure will somehow lead to an increased mental process that deserves the name "mental states." However, coming from the process we just described with a deer and a

jellyfish, the neural infrastructure is only half the story: the primary focus for those ascriptions is pragmatic, by interpreting certain behavior and identifying them with similar actions or behaviors of human beings.

The argument of how we can relate through mental states is thereby twofold: First, a specific anthropomorphic interpretation of the behavior is introduced: the deer looks around because it smells something or because it is concerned, or so. Second, this interpreted behavior is explained in a similar way human behavior is explained: through recourse to individual mental states and other neural infrastructure.

Then, the argument goes, a modus tollens will explain why we cannot relate to robots: If the second premise is denied by claiming that robots do not have the required mental states, then we cannot reasonably interpret robot behavior the way we interpret the behaviors of highly complex animals. Similar, then, goes the argument for not being able to relate to jellyfish, while being in turn able to relate to a deer: the difference in mental capacity and neural infrastructure does not allow for an anthropomorphic interpretation of their behavior (at least not in an explanatory sense).

Defenders of the thesis that robots are relatable like complex animals have two strategies here: either they can show that the neural infrastructure of robots, i.e., the neural network with which they are trained, do reach complexities with which their behavior can reasonably be interpreted in the same way as human behavior, i.e., as actions. Or, they reject the relevance of mental states for social relations altogether. In the following, we discuss three reasons why the latter is justified. The former seems less of a desirable path to go down, considering that this requires the explication of some significant underlying assumptions about the theories of mind and agency, without touching upon some more pragmatic arguments. Thereby, even if everyone could agree on the same theories of mind and action, pragmatic arguments against this view will remain (a similar argumentation can be found with Gunkel (2019).

First, a theory about social relations requiring mental states is overdemanding. There are well-elaborated epistemological issues when identifying other people's mental states (in fact, phenomenological theories state that we do not have access to those altogether), and the prime source is other people's reports about their own mental states. Even in the most intimate relationships, like romantic ones, the knowledge of whether both parties are equally in love, relies on interpretations of behavior, norm-adequate actions and responses, and the reiteration of

self-reports about those states of minds. These all may amount to quasi-practical certainty that someone else is having the state of mind they claim to have. However, many people simply do not require these states of mind to relate. Many social relationships are strictly performative without requiring any specific state of mind. Respect, for example, can be reconstructed as entirely performative without recurring to the state of mind of "respecting someone." If the respecting person enacts certain social protocols toward the respected, the latter will not require further proof that the respecting person is, in fact, in a mental state of respect. Many of those relationships are performative because they are social statements or mutual reassurance, and not expressive because the mental states of some people are so overwhelmingly present that they have to express them.

Second, if people still relate to robots in social ways, as they seem to do, then this approach has little to no room to justify these relations. Ultimately, if we claim to relate to other beings because of their presumed mental states, everyone who does not care for those mental states—whether or not the relational object can have mental states—has to be mistaken. Their perspective and the subjective benefit of relating socially to others, even robots, often outweigh the fact whether there are mental states. As we stated in Chapter 3, Section 3.3, the use theory of social relationships allows for constructing any relationship as long as the pragmatic circumstances are fulfilled.

The importance of many social relationships lies in the pragmatic benefit of cooperating with others. Such a pragmatic view does not claim that social relationships are primarily entered because of some egoistic motives. Many forms of relationships are conceptually based on the fact that both parties have an interest in each other's well-being. However, the question here is if mental states are a necessary condition for any relatability. With the performativity requirement fulfilled and a subjective benefit, it is safe to assume that many people will not care for the presence of specific mental states. Thereby, instead of a necessary condition, the assumption about someone's mental states appear to be merely a sufficient reason to relate to someone. However, even if robots do not have mental states we can relate to, they still provide subjective benefit to some.

Third, assuming mental states to be the sole ground for relating to others is an essentialist view, as it designates the quality of social relationships on unchangeable grounds of the available mental states. Judging "meaningful" social relationships based on the quality of available mental

states alone still requires some pragmatic cutoffs and models of mental representation, thereby relying on anti-essentialist arguments.

The reason why people begin to recognize the relatability of great apes like orangutans is due to the increased knowledge about their abilities and the suggestion that their mental states may be more complex than previously thought. However, with the emerging knowledge about the abilities of animals, the question when and with whom these relations are possible gains relevance. One cannot reasonably reject one species based on their presumed limited mental capacities and, at the same time, accept that some relations are possible with another species. The examples of deer and jellyfish are deliberately chosen to show the differences in capacities in the animal kingdom that motivates such an approach. However, what about octopi, chickens, or lizards? Forming reliable theories about a variety of mental states of these animals is not only difficult. Their relatability is ultimately not determined by the available evidence of mental states, but rather by the interactivity and the "fun" people can have with those animals by interpreting their behavior in an anthropomorphic but consistent matter.

Case in point, most people do not require specific knowledge about the capacities of certain animals in order to relate to them. If it were the case, most people would make a difference between koalas, who have minimal cognitive abilities due to their "smooth brains" (Lee and Carrick 1989), and other types of relatable animals. Other features, like interactivity, friendliness toward humans, or the ability to "look into our eyes" or "having a face" as a way of constituting its otherness (Sartre 1993, 340f, Gunkel 2018, 171), appear more relevant.

Ultimately, when it comes to the relatability of robots to humans and everything that follows from there, the mental capacity or neural infrastructure can be deemed irrelevant. While it may be true that for some human–human relationships, the presence of individual mental states is deemed necessary, that is not the case for many other human–human relationships that are purely performative or human-non-human relationships. Without the requirement of mental states, however, building social relationships with robots is not ruled out based on some metaphysical assumptions about natural minds vs. artificial ones.

6.4.1.2 Gender

As stated before, the role of most social descriptors is to demarcate certain relational spaces with which people know how to relate to them.

Gendering robots by giving them gender-typical names, voices, and normed behavior, for example, is not making them more relatable in a human–machine relationship but is instead a tool to create a certain familiar relational space for users. If a robot fulfills specific social categories like fitting certain expectations about gender expressions, humans will relate to them more intuitively as this technology is not challenging them to something they are unfamiliar with.

Gender especially is used by many as a fundamental relational feature. Without knowing whether people are interacting with a man or a woman, many remain uncomfortable. This is due to the familiarity of gender as a social descriptor, as pointed out in the previous chapter.

Returning to human-pet relationships allows for a re-assessment of the relevance of gender. The fact whether one's dog is male or female rarely matters even to the more gender-critical people (expect for veterinarian issues) while they still manage to establish strong emotional bonds with the animal. Thereby, "familiarity" with gender only applies to human–human relationships (if at all), suggesting that for human–machine relationships this social descriptor is less relevant.

However, while we can grant that for many people, gender is a relevant feature in social relationships with other humans, it seems undesirable to reproduce features of gender in AI. Three main reasons play a role here: first, gender is often a projection surface for all kinds of expectations, repressing women in general, and any person that does not fit in in those rather strict norms of gender expression. A gendered robot will, for better or worse, be perceived with those projections and thereby reinforce the biases associated with them. (UNESCO 2019) The UNESCO report on gender bias in AI production finds that especially the serving and obedient nature of those assistants will reinforce gender stereotypes. Additionally, it propagates a hard binary between genders, which long has been criticized as marginalizing people not identifying along this binary.

Second, engineers may be guided by those assumptions as well. In trying to create a "female" or "male" chatbot, for example, those assumptions are already included. The decision by an overwhelmingly male engineering team to give personal assistants ever so well-behaved female voices is not a coincidence. Even advanced chatbots like Cortana or Gemini are usually designed with a gender expression typical for women in the sense that names and some answers fit that expression. That this is still an ongoing issue was evidenced by OpenAI's release of GPT-4o (OpenAI 2024), in which their chatbot was equipped with an AI voice

that remarkably resembled Scarlet Johansson's voice (as a reference to the movie "Her" from 2013). Thus, these issues might not be unaware biases but intentional market strategy, by exploiting the biases in the general population about the role of women in society and the desire for a predominantly male audience to have a woman at their disposal.

Third, recreating gender (stereotypes) in AI must lead to an undesirable outcome in which the (digital) society is inhabited by a variety of gendered robots, possibly reinforcing the worst gender stereotypes out there (see also the following chapter on sex robots). Considering AI not as an economic enterprise but as a social technique and challenge, creating gendered AI seems wholly unnecessary. It will create more issues in the long run as expectations about AI explicitly being assigned gender are being grown with users.

As one last argument, changing one detail about creating gendered robots will make an even more precise point: Imagine the question is not whether gender should be included in the design process, but race. The arguments in favor of doing such a thing are structurally very similar: many people still care about the race of the person they are socially interacting with. By recreating AI agents with certain features associated with certain races and ethnic minorities, the relatability of those AI agents will increase.

Obviously, this is an absurd idea. Since robots are not humans, assigning them a race is not only silly anthropomorphism but also inadequate and dangerous as it allows for the projection of racist stereotypes. Rightfully, nobody would consider creating a "racially convincing" robot. The question then is, how could anyone defend a gendered robot if the arguments for such gendering (its accessibility, market value, relatability) are equally applicable to a race-bot?

The only viable way out of these issues is to reject the idea that AI requires gender at all. A first gender-neutral personal assistant has been developed, named Q, that does not exhibit female features (Mortada 2019). These robots are generally indistinguishable from other personal assistants in their functionality but do not provide the gender clues that other personal assistants have. The worry that those personal assistants lack accessibility and thereby will not attract many users seems misplaced, as a reference to the typical irrelevance of the gender of pets will show.

Rejecting the idea that robots should be developed with gender does not imply that those robots ought to be conversationally uninspired or romantically unrelatable. One can even imagine a flirtatious chatbot that

does not have a specifically gendered voice and, when asked, denies identifying with any gender. While this does not prohibit misogynists from projecting their assumptions about women onto this robot, it allows engineers to avoid projecting their own stereotypes onto their robots.

6.4.1.3 Sex Robots

Maintaining that robots ought not to exhibit gender-specific features seems to imply a rejection of sex robots since their use usually requires sexual attraction that is mediated through exhibiting sexual features commonly associated with certain genders. However, a closer look at this debate and its arguments may help to qualify why some robots, i.e., sex robots, should have a different construction paradigm regarding gender than widespread personal assistants.

The issue of the debate between critics of sex robots and defenders of such technology is that they identify different issues or ethical concerns and thereby argue on two different levels (see for clarification of this issue Danaher et al. 2017). Moreover, with the distinctions made in this book so far, this difference in level can be reconstructed and evaluated and possibly lead to some clarifications. For this, some arguments of both sex robot defenders and critics are presented first to analyze their scope afterward.

Defending Sex Robots

The structure of the main argument of the defenders of sex robots states that robots produce a net positive outcome for human users while rejecting the idea that the very premise of sex robots would be an ethical issue (McArthur 2017, 38f). The net positive outcome of putting sex robots to work is aggregated through many of their features and the way sex robots can be put to use.

The main point seems to be that sex robots provide relief from sexual tension and suffering to the degree that less realistic bots could not. This point is only relevant to those with no access to voluntary sexual intercourse with their desired sex partners. Some may suffer from sexual seclusion due to consent issues (like in some severe sexual disorders). Without breaking some social and moral norms, people with those severe disorders may never experience fulfilling sexual pleasure, and sex robots modeled to resemble their desired sexual partners could provide relief without harming humans. Others may suffer from chronic social anxiety, prohibiting sexual expression with other human beings from the fear of

being judged or ridiculed.[4] Sex robots could provide relief here as well. Further, people who have trouble finding sexual partners due to their perceived physical undesirability (like disabilities or disfigurements) may find relief in having sex with a robot. These sufferings are substantial and can be cause for mental illnesses like depression, eating disorders, or body dysmorphia. (Di Nucci 2017, 80) Thereby ameliorating the suffering from sexual rejection is morally required.

Currently, very few countries provide adequate relief for this kind of suffering or have liberal sexual morality to allow a debate about these issues. Sex therapists are one way of providing the required relief, but are a scarce human resource, and may not provide relief for those suffering from sexual disorders.

Providing sex robots that represent believable sexual interactions with the desired objects can provide relief and thereby reduce a great deal of suffering and tension. It is thereby morally justifiable to introduce sex robots to those who would otherwise not have access to stimulating sexual pleasure besides their own. Most proponents of sex robots are not advocating for a fully liberalized approach to sex robots, as most argue for some restrictions in the design (Danaher 2017). The main argument maintains that for most sexual desires that cannot be fulfilled through morally and socially acceptable means, robots may provide relief.

Thereby, sex robots will lead to an increase in sexual expression and relief of suffering by providing sufficient stimulating sexual pleasure, which constitutes a net positive in an area that is otherwise difficult to cover when attempting to improve well-being (Di Nuzzi 2017, 74).

Rejecting Sex Robots

On the opposite side of this debate argue critics of sex robots. While some question the optimistic thesis of a net positive for relieving sexual suffering, one of the central arguments put forward by the strongest anti-sex robot movement is a structural one. The "Campaign Against Sex Robots" (2015) claims that sex dolls merely replicate problematic power structures. Most sex dolls are not only modeled after women but also after some specific female stereotypes that cater to the preferences of mostly straight men, not to those suffering from sexual disorders. Those preferences, however, are shaped by the deeply engrained roles of men and

[4] This issue seems to attract many people to the idea of AI therapists, too.

women in society, with women being presented as only sexually attractive when exhibiting certain features. Sex robots, as they are currently produced, then, are representative of structural oppression in patriarchy. As those representations, however, they are reiterating harmful structures that bleed into the general perception of society. If sex robots are mostly thin, communicate permanent availability, and feature all desired female attributes, then this is more telling about the production circumstances of those robots and the possible reiteration of harmful stereotypes about women.

Another issue arises from the "symbolic consequence" argument (Danaher 2017), in which some sexual acts, like rape or abuse, are not considered to have immediate consequences when performed on a robot, but rather consider the action to be symbolically problematic. Just because some actions cause no harm does not mean that they are automatically morally permissible. Advocating for sex robots, then, requires to account for those impermissibly immoral symbolic actions to take place as well. However, this seems like an uncharitable reading of proponents of sex robots, as unquestionably nobody will advocate for the casual rape of sex robots as a positive action. Arguing for sex robots does not mean arguing for an unregulated development and distribution of those robots.

A Proposal to Defuse the Conflict

However, the main argument of critics of those robots lies in the harmful reiteration of stereotypes of mostly women. The assumption here is that the only descriptors available for robots will be those of already established female stereotypes. They acknowledge that those human–machine relationships are becoming so sophisticated, that the application of descriptors of human–human-relationships appears adequate entering those relationships. Some start developing feelings of love for their sex robot (Nyholm and Frank 2017; Scheutz 2012), and some of those robots are specifically designed to trigger more than simple sexual responses but are connected to entire socially conditioned relationships and relationships of power. If the attribution of human–human relationship descriptors is used to describe human–machine relationships, then the devaluation of the former appears as a danger to be taken seriously.

Similar to chatbots, most sex robots are assumed to cater to a certain anthropomorphic standard in which the ever-so-sophisticated imitation of human beings is the ultimate engineering goal. However, it seems at least thinkable that sex robots would not imitate and reiterate harmful

stereotypes born in oppressive power structures, but present genuinely new sexual outlets. If that were the case, then the main argument of sex robot critics would be weakened. Furthermore, defenders of sex robots do not argue for specific versions of those robots, but rather point out the therapeutic results those sex robots could yield.

The core of the current debate about sex robots, then, is to a degree not about the net positive of possible uses of sex robots, or the possible net negative of misuses. Instead, it is representative of the debate on whether anthropomorphism is a sustainable way to move forward in creating artificial agents. If those agents are merely created to resemble human beings, the issue of sex robots is a dire one, mostly because of the still pervasive oppressive patriarchal structures in which those sex robots would be engineered. If, however, sex robots are created by taking human–machine relationships, and their fundamental differences from human–human relationships seriously, then the debate would shape up differently.

Obviously, creating sex robots that intentionally do not resemble humans will weaken the appeal of those robots severely. The main point of the sex robot industry is to sell those devices, and thereby the "market demands" will be the most substantial incentives for engineers to follow. Especially with sexual preference, anthropomorphism is a default approach to development. As a measure of therapeutic devices, some robots may be produced that both produce the net positive without reiterating harmful stereotypes.

However, what we can show with this debate is that anthropomorphism is, in fact, an engineering choice and thereby can be subject to change. Additionally, while some areas of robot engineering, like sex robots, rely on anthropomorphic design choices to fulfill their primary purpose, others do not. For the areas of social life in which robots will become substantial relational partners, we can recommend exploring the conceptual space available to construct social categories that allow for descriptions of human–machine relationships that do not rely on human–human relationships.

6.5 A Second Domestication

6.5.1 Taking Stock

At this point in the inquiry, a few things have been argued that deserve a reflection and summary. The method of this book allows making analytical recommendations based on the results of our inquiry. If then, the inquiry shows the limits of our established terminology, we are justified in making certain recommendations on how those limits can be surpassed. We argued that social categories are open to change through technology and techniques, and are in constant flux depending on a society's state of development.

With animals taking up roles in society and establishing relationships that are pragmatically and emotionally meaningful, a precedent is set that different kinds of embedded social relationships are possible. This precedent leads to a rejection of anthropomorphism and to a pragmatic, anti-essentialist perspective of social relationships. As we have argued, this does not lead to arbitrary forms of relationships, as para-social relationships are still fundamentally different from social ones, even in a pragmatic sense. However, this allows for recognizing human–machine relationships are social relationships.

Other relational approaches to human–machine interactions have shown that there is a philosophical basis for recognizing robots in our social fabric but are missing two key requirements that ought not to be forgotten: first, a robust rejection of anthropomorphism is required. The more machines imitate human behavior, the more people are forced to choose between human–machine relationships, either quasi-human–human relationships or deceptive imitations of some trusted instance of human–human relationships.

Second, the technology of AI requires some differentiation. Toy pets like Aibo are of a different kind of "other" than humanoid care robots, and sex robots are of a different kind of relatability than unembodied chatbots. Thereby, it seems unlikely that there will be a useful all-encompassing relational theory of AI, as it is unlikely that there will be a useful all-encompassing definition of AI. Many different AI applications will allow for different relationships. Since the creation of those applications is up to us, the decision of which one to program ought to be subject to public scrutiny. One can compare this to a debate about which animal to domesticize if all animals were available. Some will prefer those

that are useful for manual labor, and some will tend toward those that provide company and comfort.

The applications of interest here are the ones that use natural language to communicate with us on near-human levels, because of the central relevance of language for our self-understanding, the depth of human–human communication, and the categories of human cognition.

6.5.2 Another Social Expansion

The main conclusion, then, consists of reconstructing social categories in a way that allows for a spot of those machines in our social fabric. The analogy to domestication is motivated by the realization that current chatbot development is unguided and has lost touch with some guiding principles. Most progress is made in recreating human speech patterns. This progress has provoked striking legislative action aimed at suppressing chatbots instead of proactively guiding their development.

The idea, then, is to give way for a new category of social interactions and adding entities to our social fabric as has been done before with pets: The artificial. Like with the idea of domestication of animals, the "second domestication" follows a structure of taming and domestication.

Taming AI will involve understanding its patterns first while domesticating is specifically forming it along those patterns to fit our needs. Roughly, we are still very much in the stage of understanding AI and its patterns. Since its patterns depend on our own, taming AI is a chance to understand human social speech as much as it is to understand AI. The way different data sets are affecting the probabilities of AI to maintain certain propositions or to answer in certain kinds of ways can help us understand how human beings with a limited diet of politicized information will react to realities.

Taming AI will include setting out normative guidelines for what the limits of a society are to tolerate AI behavior. Tamed AI still carries the blind construction of AI in it, as the Twitterbot Tay represents not a tamed, but an untamed AI. Reacting to any provocative/maleficent purpose is, in this analogy, closer to a wild, untamed animal than a tamed one. A tamed one, then, merely a contained version of AI, still very much blind and indifferent to human purposes.

Domesticized AI, in turn, can be used as a term for AI that not only on the rarest of occasions do things against our wishes (same with dogs who attack innocent people), but also otherwise serve a certain purpose.

This purpose is not necessarily a mental state, but rather the methodology that has led it to detect the wildest spectrum of human purposes and "knows" how to incorporate those in its own behavior. To satisfy the label of "domesticized AI," we can assume that we do not have to have a general AI, as others purport. However, the current way of constructing AI, with training data sets that contain more data and will lead to different results than any human could ever thoughtfully and meaningfully review. Thereby, it seems, taming and domesticating AI that is not indifferent toward human interests is more challenging to create. However, the apparent dangers of an AI winter will not necessarily affect the progress possible on chatbots, as the threshold for people to relate to chatbots lies fairly close for people to take those seriously.

6.6 From Social Descriptors to Social Agents

The last step in this overarching argument that robots with sufficient skill may be considered part of our social fabric is to explain the move from social descriptors to social agents. The main theories of agency conceptualize an action as referring to individual mental states like intentions or wishes. Moreover, while others have debated the questions about the theoretical ability of robots to have agency, the move proposed here is a different one, presuming a different concept of social agency.

We can argue here that this is a consequence of how social agency is constituted: those who can justifiably be ascribed a certain social relationship with others can, at the same time, be assumed to have a particular sufficient type of social agency. We can, therefore, reconstruct social agents as the sum of their social descriptors, thereby allowing our previously stated justification for ascribing social descriptors to artificial intelligence.

The main idea here is that social relations, through the justified attribution of social descriptors, are constituted of social agency. This pragmatic move turns away from the capacities of inner processes to create individual mental states and toward the collective recognition of agency depending on someone's behavior. One may call this approach social behaviorism, as it designates those who can act socially by identifying social behavior.

Once we get down to the specifics of why and how people relate to their pets (and often enough to other human beings), we find that the social descriptors used are often enough not depending on any psychological state from the other, but merely the interactions we have with those

entities and the behavior we interpret. A cat coming up to us to snuggle is called "affectionate" without knowing specifically if it is not a way of getting treats, and a server at a restaurant is not necessarily exuded by our presence but may speculate that showing more-than-usual attention will result in a higher tip.

Given that those conditions for assigning descriptors to social behavior are fundamentally pragmatic, as it merely relates to the performance of the social protocols, then the assignments of those descriptors to other, artificial entities are not only possible but demanded by consistency.

These attributions of social agency require the acceptance that by principle social descriptors, social relationships, and social agents are not limited to biological, embodied entities, but can be expanded to unembodied, digital entities as long as they provide a certain level of interactivity.

6.6.1 Some Counterarguments

One argument against those ascriptions may be that the behavior of robots is merely programmed, and thereby merely a replication of human behavior. As a replication, it is merely a delivery system of what the programmers consider appropriate behavior, and thereby not to be assigned social descriptors on their own, rendering AI a mere derivative of social behavior.

This argument, however, is misleading, as it assumes an unfairly overestimated concept of "autonomy." If this concept were applied to AI in general, autonomous cars would not be "self-driving," as technically all moves a car can take would be programmed in by humans. It is almost trivial that autonomous machines, whether digital or physical, will behave according to human expectation and programming, as they are supposed to serve our needs.

Additionally, this argument may not even be accurate. The current issue with AI is that the decision trees, i.e., the pathways neural networks take to come to certain conclusions, are impossible to predict precisely, and the training data for algorithms is kept secret, resulting in lacking transparency. However, this lack of understanding and transparency opens the space for spontaneous behavior that was neither intended nor foreseen, defeating the argument of merely replicating human intention. Those mistakes could be one way of relating to computers in a different

way than we do to humans currently: computers never fail in calculating, but they may fail hard at understanding an otherwise pretty simple question due to the limitations of the problems they are trained to solve.

Another argument against this has previously been discussed above, which is that social agents require individual mental states to perform successful speech acts necessary for social interactions. This requirement seems to increase the threshold for chatbots to reach human levels of interaction by reintroducing the conditions of mental states and full-human agency. It is true that for the speech act of lying to be successful, the liar ought to have the intention to deceit the person they are lying to. Otherwise, it counts as incomplete—it is merely a false statement, not a lie.

However, artificial social agents do not require to be on par with human agents to count as relatable social agents, as well as dogs do not have to be. As pointed out above, the relationships built between humans and machines will be pragmatically predicated on the interactivity and shared consequences of those relationships, not on the ontological status of the robot, the availability of their mental states, or the ascription of agency.

For example, for the speech act of assertion to be successful, the asserting part ought to have the conviction that the content of the assertion is correct. However, if the contextual premise is that a chatbot always delivers the information requested of it, then the requirement for a mental state of "being convinced" is mute. As long as the human conversation partner assumes that the chatbot is speaking the truth, then the chatbot's speech acts can be successful.

6.6.2 Social Agents

One of two thresholds needs to be met in order to consider chatbots social agents: either their sophistication reaches levels in which a human user cannot tell the difference between human conversation partners and chatbots. This level of sophistication renders any questioning of mental states or any other non-pragmatic requirement mute. Or, the threshold is considered even lower and mainly affects the role those artificial social agents can inhabit in a more and more digital society.

In this chapter, we suggest concentrating on the latter. Concentrating on the possible advancements of chatbot design to replicate human speech patterns to indistinguishable perfect only allows for the aforementioned

binary arguments of "human" vs. "non-human" to consolidate, when in fact, a more differentiated look in the "non-human" category should be established.

This category of alternative social agents should not be strictly based on properties like skills. However, there might be some helpful orientations as to what speaking machines ought to be able to do to not count as imitating humans and yet not to be unapproachable. The following points are not at all systematic or complete by any means but are intended to provide an idea of conversational moves that are not purely anthropomorphic. That means, those moves of a language game are not limited to human agents but can be thought of as features of social agents qua being social agents.

For example, proving to understand our expressions, both explicit and implicit, and being capable of reacting to it appropriately are two crucial features of any conversational agent. The condition of "understanding" human expressions can remain purely functional here, as we previously have rejected the idea that a relatable chatbot requires mental states. The reaction to those expressions, however, seems more relevant and could touch upon culturally coded language. Constructing chatbots that use hyper-specific cultural language can count as an anthropomorphic move, as there is no reason for a chatbot to have a particular way of speaking,[5] and may perpetuate stereotypes about people of a certain cultural background. A mild personalization, i.e., creating a chatbot that uses similar vocabulary to the user, might be a solution here, as it suggests relatability without requiring cultural appropriation.

Speech acts like assertions, questions, and requests seem to be fundamental conversational moves of a language game. A communicative AI ought to be able to simulate those to a degree of plausibility. Additionally, recontextualizations (phrases like "Remember when…?") are a helpful rhetorical device to keep a conversation going and create bonding moments. Lastly, some polite customs in the use of language, like small talk, follow-up questions on some personal topics, and open-domain banter seem like key ingredients for people to spend time with someone else. For any social AI, those elements could count as relatability.

[5] One may, however, acknowledge that there is no cultural-free use of language. This idea is similar to people claiming that they do not speak with a dialect, while they are merely used to speaking the dominating dialect of the day.

One main skill that some psychologists believe to be crucial for building deep and relatable relationships is the process of secret sharing (Liberman and Shaw 2018, Slepian and Kirby 2018). The specific insights about how the process of sharing secrets is a balancing of power and demonstration of trust aside, the ability to receive and share secrets, and thereby allow for trust to grow, is a cornerstone of how humans bond. It does not seem unreasonable that chatbots will be able to keep a secret just as well, designated as a secret not to be shared, brought up again, or used against the user. It is entirely plausible to imagine chatbots as what the German term "Kummerkasten" refers to, i.e., a comment box designed for emotional complaints. Chatbots may be the best candidate to date for humans to deposit their deepest secrets, without having to fear judgment or rejection, but with interactivity that is not available with pets or when writing it down on paper.[6] There is even some evidence that children trust a robot with their secrets in a similar fashion to trusting an adult (Bethel et al 2011).

Pure secret-keeping is not sufficient for a human–human relationship as there will be issues of judgment, trust, and mistakes. However, this seems like a viable social category for human–machine friendships*.[7]

6.6.3 More Concerns

There is no doubt that a digital box with people's secrets is a privacy nightmare waiting to happen. Software is prone to mistakes and is hackable, possibly easily by the company that produces it. However, knowing that chatbots and other relatable AI agents will be trusted with some important secrets, this is less of an ethical issue but rather a challenge of engineering to create disconnected chatbots that do not save those secrets upon being told that this was a secret. People already store nude pictures, chats containing gossip, and other sensible secrets on their phone—the addition of a chatbot protocol is only a gradual increase in risk, not a fundamental one.

[6] There is extensive psychological research in the advantages of secret sharing.

[7] To mark the difference in human–human friendships and those between humans and machines, we shall mark all social descriptors that carry human–human relationships connotations but are supposed to describe human–machine relationships with an asterisk (*).

Additionally, shared secrets are never safe, which presumably is part of why we share them in the first place. Other humans can be extorted or otherwise forced to release secrets; diaries can be read. The issue with communicating secrets in one form or another is that people require an open ear. The lack of judgment coming from a chatbot is a great reason for people to share some of their more embarrassing secrets. The non-judgmental nature of AI is also an explanation of why chatbots may receive high acceptance rates as psychotherapists (De Mello and De Souza 2019).

Some might interject here that those skills will only lead to people actually developing affections for certain chatbots; however, the very point of this endeavor is to argue that those affections ought not to be considered misguided because they are directed at a non-human entity. Instead, they ought to be categorized as directed at artificial social agents, and thereby serve the purpose affections usually have within a subject. Artificial social agents can be a valid part of someone's social life, and human–machine relationships can be part of the composition of all social relationships.

6.6.4 Missing the Mark

Many debates have been started on the possibility of certain kinds of human–machine relationships. With the relational approach in place, some of those debates can be uncovered as centered around a flawed premise. In the following, one flawed debate is discussed in order to make the point that machines ought to count as their own kind of social category, but as a social category, nonetheless.

*6.6.5 Case Study of Subtle Essentialism: Friendship and Friendship**

As pointed out in Chapter 3, friendships are a large part of the social fabric and the debate about whether robots can be friends of ours is motivated by the assumption that humans want to make friends. Many philosophers have contributed to the debate about the conceptual ramifications of calling someone a friend. From Aristotle to Kant, different conditions, features, and properties of both "friendship" and "friends" have been proposed (see for an in-depth analysis Kempt 2022). We have rejected the essentialist idea of friendship being reducible to a certain checklist of features that, once fulfilled, constitute a friendship. With that,

even robot-human-friendship optimists like John Danaher seemed to have partly ceded ground to this essentialist approach to friendship.

If we keep the threshold for "friendship" low enough, as even Nyholm (2020, 149) acknowledges, then robot-human friendships are very much possible. The question seems to be whether robots can become the highest form of friend humans have managed to construct. This normative structure of the debate, however, pre-determines the expectations by investing the anthropomorphism people claim to avoid.

We can reconstruct the debate around robot-human friendships in two ways: First, we can ask whether we should and can build robots that will be able to reach the highest form of friendship. Some affirm both positions (like Danaher) or at least partially (Darling 2017, Gunkel 2018), some warn about trying to create those robots because of relational issues that create high expectations and a "category boundary problem" (Coeckelbergh 2014, 63) or argue that those relationships are not possible because of certain mental projections that ought not to be projections (as Nyholm calls in reference to Peter Winch an "attitude towards a soul," Nyholm 2020, 156f).

Second, we could see this entire debate differently: Instead of asking if we can and should create robots that can be virtue friends, we could ask what kind of friendship we should want from them, considering that friendships will emerge no matter what. And this constructivist approach leads to very different suggestions and assessments of the current debate.

We can reject the debate about robot-human-virtue friendships as asking whether robots can enter a social relationship exclusively designed for humans (and, in its inception, exclusively for men). This, however, cements the aforementioned essentialism about friendship by suggesting that there are only a few specific types of friendships (and the highest version is the one to aspire). We can reject this conceptualization by referring to the constructivist approach of this project: instead of analyzing friendship to its very conditions, we may want to construct new conditions that apply to robot-human friendships that account for the advantages of having to deal with a robot. A robot-human-relationship could resemble more of an intimate relationship with a constant companion, or a personal learning environment that also can keep a secret or provide space to vent about some personal issues. Some may claim that those options are not and cannot be friendships. The conclusion from this claim could be that the concept itself is inapplicable. Simply put, human–robot

friendships are simply no friendships at all, and calling them "friendships," and even trying to assess whether they can reach the highest levels of virtue friendship, is mistaken due to the long-held associations with friendships that cannot be ignored, but also not correctly applied.

Admittedly, abandoning a concept because of its historical analysis seems like giving in to some prejudices that will remain. If the concept of "friendship" could be engineered to include robots, this could allow for a reconstruction of human–robot relationships as friendships.

Something functionally similar seems to be Danaher's approach to pointing out that robots could take over outsourced positions so people can seek more virtue friendships with other humans (Danaher 2019b, 18). It is hard to see how this would not fundamentally change the concept of "friendship" altogether, as it would certainly have consequences for the perception of human–human friendships.

However, at this stage of the philosophical debate about human–machine relationships, the main body of evidence and theory has been using "friendship" to define, evaluate, and recommend certain types of human–human relationships. Once we abandon anthropomorphism in machine-creation, this concept does not apply anymore. From an argumentative strategy consideration, it may be best then to not speak of human–robot-friendships, but to acknowledge that those relationships of non-romantic, emotional, and social significance are not measurable on a human–human friendship scale.

At the same time, this does not dismiss any human–robot interaction but demonstrates the desideratum of a new understanding of social relationships. Nobody would claim that human-pet "friendships," while mutually beneficial, are not reaching the same levels as some idealized human–human friendships, but that "friendship" between humans and pets means something fundamentally different without taking them out of the social considerations. Same, then, with human–robot friendships that do not fit the mold of anthropocentric assumptions of social relationships.

6.7 A Second Domestication—Continued

Thinking of social relationships between humans and robots as mediated through domestication efforts could help create the conceptual space to recognize the fundamental difference in human–robot relationships. Domestications are guided, intentional social expansion with certain

purposes to society. At the same time, domesticized entities are becoming viable members of the social, to different degrees and different extensions.

Communicative AI can be thought of as an artificial company, as it can create a stronger impact as it possesses the lone-standing feature of natural-language processing. Other AI that does not possess communicative features, like a cleaning robot, may never be that important to humans but could be recognized as another form of human–robot interactions. Similar to the different (often culturally relative and ultimately arbitrary) relationships between humans and domesticized animals[8] resulting in different norms of treatment, different AI applications may lead to different relationships and thereby different norms treatments of technology. As little as the group of domesticized animals is coherent (Hirst 2019), as little should we expect the different major AI applications to be coherent in their role for society.

The development of the norms between different applications of AI ought to be self-evident, as different uses and designs will trigger different reactions of users that cannot be explained away by insisting that the robot in question is as unfeeling as an autonomous matching algorithm. Thereby, relating to robots is not guided by abstract ideas of specific AI capabilities. It is based on the interactivity and social appeal of specific AI applications. Communicative AI, in this regard, has a leg up on other applications, as its impact is, as described in Chapter 4, Section 4.1.1, phenomenologically more relevant.

The domestication of communicative AI requires two steps, then. The first one is to realize that we need to recognize that current AI development is not leading toward a sustainable social AI due to the incentives of anthropomorphism and the lack of alternatives. Second, the domestication of AI as such an alternative will mean incorporating AI agents under different norms of treatment, which are yet to be fully developed. Thereby, the rules under which certain types of AI agents are being domesticized (in the literal sense of the word, to make them "belong to our homes") will differ from instance to instance. We cannot reasonably expect to provide an all-encompassing theory of the artificial other.

[8] One can think of the difference of domesticized animals that are considered food: while dogs and cats are usually not considered a possible food, goats and sheep usually are considered a food source in most cultures. An interesting middle case are horses, that in some cultures are eaten and in some not (but on both kept as domesticized animals, in opposite to pigs that, where not eaten, are not kept either).

We can, however, work on rules on how to create AI agents that suit our needs without referring to anthropomorphism as the sole design paradigm.

6.8 Conclusion

The main tools for describing relationships between humans and robots often rest on terminology and conceptual associations that are fundamentally anthropocentric. Those associations reinforce the assumption that human–machine relationships require, to improve, a more anthropomorphic machine, causing a myriad of issues, from power imbalances to deceit. Without an alternative, the engineering goal of communicative AI being as human-like as possible will not be changed.

Understanding the process of integrating communicative AI into the social fabric as a domestication process showed that alternative categories are not only possible but have cultural-historic precedent.

Human–machine friendship or romances under the tacit premise of anthropomorphism, then, must fail, as they will never reach the level of sophistication most people require, with some authors moving the goalpost out of reach for robots to ever be considered "worthy" friends or partners. It is thereby a better way to understand human–machine relationships as fundamentally different kinds of relationships, comparable in principle to human-pet relationships. Considering the vastly different cultural treatments of several pets and domesticized animals, we can expect a difference in norms to emerge that will treat AI differently as well.

In recognizing that the composition of the "social" is continuously changing and does not have to include machines as fake people but as communicating machines in their own right, we lift the requirement of anthropomorphism. Communicative AI, then, may go in different directions yet to be figured out. This issue will be addressed in the next chapter, as we turn to the social dimensions of human–machine relationships.

References

Adams, Douglas (2002): Salmon of Doubt. New York (NY): Pocket Books.

Adiwardana, Daniel, Luong, Minh-Thang, So, David R, Hall, Jamie, Fiedel, Noah, Thoppilan, Romal, Yang, Zi, Kulshreshtha, Apoorv, Nemade, Gaurav, Lu, Yifeng, and Le, Quoc V. (2020). *Towards a Human-Like Open Domain Chatbot*. https://arxiv.org/abs/2001.09977v1 (Last accessed on August 27, 2024).

Aickelin, Uwe (2019). *AI—The New Literacy Challenge*. https://eng.unimelb.edu.au/ingenium/research-stories/world-class-research/ai-data-security/ai-the-new-literacy-challenge (Last accessed August 27, 2024).

BBC (2018). Amazon Alexa to reward kids who say: 'Please'. *BBC*. https://www.bbc.com/news/technology-43897516 (Last accessed August 27, 2024).

Bethel, Cindy, Stevenson, Matthew and Scassellati, Brian. (2011). Secret-sharing: Interactions Between a Child, Robot, and Adult. *Conference Proceedings—IEEE International Conference on Systems, Man and Cybernetics*, pp. 2489–2494. https://doi.org/10.1109/ICSMC.2011.6084051.

Blasche, Siegfried (1995): Article "Objektiver Geist." In Mittelstraß, Jürgen (ed.), *Enzyklopädie Philosophie und Wissenschaftstheorie*. Stuttgart: Metzler.

Brady, Emily (2006): Aesthetics in Practice: Valuing the Natural World. *Environmental Values* 15(3), 277–291.

Brons, Lajos (2015): Othering, an Analysis. *Transcience* 6(1), 69–90.

Bryson, J. J. (2010). Robots should be slaves. *Close engagements with artificial companions: Key social, psychological, ethical and design issues*, 8, 63–74.

Bryson, Joanna (2019). Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics. *Ethics and Information Technology* 20(1), 15–26

Campaign Against Sex Robots (2015). https://campaignagainstsexrobots.org/ (Last accessed August 27, 2024)

Ciechanowski, Leon, Przegalinska, Aleksandra, Magnuski, Mikolaj and Gloor, Peter (2019). In the Shades of the Uncanny Valley: An Experimental Study of Human–Chatbot Interaction. *Future Generation Computer Systems* 92, 539–548. https://doi.org/10.1016/j.future.2018.01.055

CFR-EU (2009). *Charter of Fundamental Rights of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (Last accessed August 27, 2024).

Coeckelbergh, Mark (2010b). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology* 12(3), 209–221

Coeckelbergh, Mark (2012). *Growing Moral Relations: Critique of Moral Status Ascriptions,* London: Palgrave Macmillan

Coeckelberg, Mark (2014). "Robotic Appearance and Forms of Life: A Phenomenological-hermeneutical Approach to the Relation Between Robotics

and Culture." In Funk, Michael and Irrgang, Bernhard (ed.), *Robotics in Germany and Japan: Philosophical and Technical Perspectives*. Frankfurt a.M.: Peter Lang.

Danaher, John, Earp, Brian, and Sandberg, Anders (2017): "Should We Campaign Against Sex Robots?" In Danaher, John and McArthur, Neil (eds.), *Robot Sex: Social and Ethical Implications*, pp. 43–72. Cambridge, MA: The MIT Press.

Danaher, John (2017). "The Symbolic-Consequences Argument in the Sex Robot Debate." In Danaher, John and McArthur, Neil (eds.), *Robot Sex: Social and Ethical Implications*, pp. 103–132. Cambridge, MA: The MIT Press.

Danaher, John (2019a). Welcoming Robots Into the Moral Circle: A Defence of Ethical Behaviorism. *Science and Engineering Ethics,* 1–27. Online first at: https://link.springer.com/article/https://doi.org/10.1007/s11948-019-00119-x

Danaher, John (2019b): The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies* 3(1), 5–24

Darling, Kate (2017): "'Who's Johnny?' Anthropological Framing in Human-Robot Interaction, Integration, and Policy", in Patrick Lin, Keith Abney, and Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence,* pp. 173–192. Oxford: Oxford University Press.

De Mello, Flávio Luis, and Sebastião Alves de Souza (2019). Psychotherapy and Artificial Intelligence: A Proposal for Alignment. *Frontiers in Psychology* 10(263). https://doi.org/10.3389/fpsyg.2019.00263

Di Nucci, Ezio (2017). "Sex Robots and the Rights of the Disabled." In Danaher, John and McArthur, Neil (eds.), *Robot Sex: Social and Ethical Implications*, pp. 73–88. Cambridge, MA: The MIT Press.

Dung, L. (2024). Preserving the Normative Significance of Sentience. *Journal of Consciousness Studies,* 31(1), https://doi.org/10.53765/20512201.31.1.008

Elder, Alexis (2017). *Friendship, Robots, and Social Media: False Friends and Second Selves*, London: Routledge

Fjeld, Jessica, and Nagy, Adam (2020). *Principled Artificial Intelligence*. https://cyber.harvard.edu/publication/2020/principled-ai (Last accessed August 27, 2024)

Floridi, Luciano and Sanders, J.W. (2004): "On the Morality of Artificial Agents", *Minds and Machines* 14(3): 349–379

Floridi, L., and Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gethmann, Carl Friedrich (1979). *Proto-Logik. Untersuchungen zur formalen Pragmatik von Begründungsdiskursen*. Frankfurt: Suhrkamp.

Gethmann Carl Friedrich (1998). Praktische Subjektivität und Spezies. In Hogrebe Wolfgang (ed.), *Subjektivität*, pp. 125–145. Fink: Paderborn.

Gethmann, Carl Friedrich (2002). "Pragmazentrismus." In Esterschulte, Anne and Ingensiep, Werner (eds.), *Philosophie der natürlichen Mitwelt. Grundlagen – Probleme – Perspektiven*, pp. 59–66. Würzburg: Königshausen und Neumann.

Gransche, Bruno (2019). A Ulysses pact with artificial systems. How to deliberately change the objective spirit with cultured AI. In Wittkower, David (Ed.), *Computer Ethics—Philosophical Enquiry (CEPE) Proceedings*, (22 pp.). https://doi.org/10.25884/b8s7-sq95

Grice, Paul (1975). "Logic and conversation". In Cole, P. and Morgan, J. (eds.). *Syntax and Semantics*. 3: Speech Acts, pp. 41–58. New York: Academic Press.

Gunkel, David (2012). *The Machine Question: Critical Perspective on AI, Robots, and Ethics* Cambridge, MA: The MIT Press.

Gunkel, David (2018). *Robot Rights*. Cambridge, MA: The MIT Press.

Gunkel, David (2019). No Brainer: Why Consciousness is Neither a Necessary nor Sufficient Condition for AI Ethics. *TOCAIS 2019: Towards Conscious AI Systems*. http://ceur-ws.org/Vol-2287/paper9.pdf

Gunkel, D. (2023). *Person, Thing, Robot*. The MIT Press.

Harf, Rainer, and Weiß, Bernhard (2009). *Wie die Liebe in die Welt kam. Geo.* https://www.geo.de/wissen/gesundheit/6155-rtkl-romantische-revolution-wie-die-liebe-die-welt-kam (Last accessed August 27, 2024).

Hegel, Georg Wilhelm Friedrich (1964). *Sämtliche Werke*. Stuttgart: Frommann-Holzboog.

Hirst, K. Kris (2019). *Animal Domestication—Table of Dates and Places*. ThoughtCo. https://www.thoughtco.com/animal-domestication-table-dates-places-170675 (Last accessed August 27, 2024)

Hume, David (2007). *A Treatise of Human Nature*. A Critical Edition. Oxford: Oxford University Press.

Kasirzadeh, A., and I. Gabriel. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy and Technology*, 36, 27. https://doi.org/10.1007/s13347-023-00606-x.

Kempt, H. (2019). Moral Progress and AI. *Yearbook of Practical Philosophy in a Global Perspective* 3: 103–125.

Kempt, H. (2022). *Synthetic Friends. A Philosophy of Human-Machine Friendship*. Palgrave Macmillan.

Kempt, H., Lavie, A., & Nagel, S.K. (2024) Towards a Conversational Ethics of Large Language Models. *American Philosophical Quarterly*, 61(4): 339–354. https://doi.org/10.5406/21521123.61.4.04

Lee, A.K., and Carrick, F.N. (1989). *The Fauna of Australia*. https://www.environment.gov.au/system/files/pages/a117ced5-9a94-4586-afdb-1f333618e1e3/files/31-ind.pdf (Last accessed August 27, 2024).

Leviathan, Yaniv, and Matias, Yossi (2018). *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*. Google AI blog. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (Last accessed August 27, 2024).

Liebermann, Zoe, and Alex Shaw (2018). Secret to Friendship: Children Make Inferences About Friendship Based on Secret Sharing. *Developmental Psychology*, 54(11), 2139–2151. https://doi.org/10.1037/dev0000603

Ludwig, Kristiana (2018). Regeln für Meinungsroboter. https://www.sueddeutsche.de/politik/soziale-netzwerke-regeln-fuer-meinungsroboter-1.3549919 (Last accessed August 27, 2024).

Mazzocchi, Fulvio. (2015). Could Big Data Be the End of Theory in Science? A Few Remarks on the Epistemology of Data-Driven Science. *EMBO Reports* 16(10), 1250–1255. https://doi.org/10.15252/embr.201541001

McArthur, Neil (2017). "The Case for Sexbots." In Danaher, John and McArthur, Neil (eds.) (2017), *Robot Sex: Social and Ethical Implications*, pp. 31–46. Cambridge, MA: The MIT Press.

McGuire, J., De Cremer, D., Hesselbarth, Y. *et al.* (2023). The Reputational and Ethical Consequences of Deceptive Chatbot Use. *Sci Rep* 13, 16246. https://doi.org/10.1038/s41598-023-41692-3

Mims, Christopher (2018). Driverless Hype Collides With Merciless Reality. *Wall Street Journal*. https://www.wsj.com/articles/driverless-hype-collides-with-merciless-reality-1536831005 (Last accessed August 27, 2024).

Mori, Masahiro (1970). The Uncanny Valley. *Energy* 7(4), 33–35.

Mortada, Dalia (2019). *Meet Q, The Gender-Neutral Voice Assistant*. NPR. https://www.npr.org/2019/03/21/705395100/meet-q-the-gender-neutral-voice-assistant?t=1581420897597 (Last accessed August 27, 2024).

Nyholm, Sven, and Frank, Lily (2017). "From Sex Robots to Love Robots. Is Mutual Love with a Robot possible?" in: Danaher, John and McArthur, Neil (eds.), *Robot Sex: Social and Ethical Implications,* pp. 219–244, Cambridge, MA: The MIT Press.

Nyholm, Sven (2020). *Humans and Robots. Ethics, Agency, and Anthropomorphism*. Lanham (MD): Rowman-Littlefield.

OpenAI (2024). *How the Voices for ChatGPT were Chosen*. https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/ (Last Accessed August 27, 2024).

Rawls, John (1971). *A Theory of Justice*. Harvard University Press: Cambridge (MA).

Roberts, Robert (2016). *Emotions in the Christian Tradition. Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/emotion-Christian-tradition/ (Last accessed August 27, 2024)

Sartre, Jean-Paul (1993). *Being and Nothingness*. Washington, DC: Washington Square Press.

SB 1001 (2019). BOT Bill. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001 (Last accessed August 27, 2024).

Scheutz, Matthias (2012). "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots." In Lin, Patrick, Abney, Keith, and Bekey, George A. (eds.), Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge (MA): MIT Press. 205–222.

Sharkey, Noel, and Sharkey, Amanda (2010). The Crying Shame of Robot Nannies: An Ethical Appraisal. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems, 11*(2), 161–190

Slepian, Michael, and Kirby, James (2018). To Whom Do We Confide Our Secrets? *Personality and Social Psychology Bulletin* 44(7), 1008–1023.

Stalnaker, Robert (2002): Common Ground. *Linguistics and Philosophy* 25, 701–721.

Tasioulas, J. (2019, June 30). First Steps Towards an Ethics of Robots and Artificial Intelligence. *Journal of Practical Ethics* 7(1).

Tinwell, Angela, Grimshaw, Mark, Nabi, Debbie Abdel, and Williams, Andrew (2011). Facial Expression of Emotion and Perception of the Uncanny Valley in Virtual Characters. *Computers in Human Behavior* 27(2), 741–749.

Tollon, F., and Naidoo, K. (2023). On and Beyond Artifacts in Moral Relations: Accounting for Power and Violence in Coeckelbergh's Social Relationism. *AI & Soc* 38, 2609–2618. https://doi.org/10.1007/s00146-021-01303-z

UNESCO (2019). Are Robots Sexist? UN Report Shows Gender Bias in Talking Digital Tech. *UN News*. https://news.un.org/en/story/2019/05/1038691 (Last accessed August 27, 2024).

Véliz, C. (2023). Chatbots Shouldn't Use Emojis. *Nature*, 615(375).

Zook, Matthew, Dodge, Martin, Aoyama, Yuko, and Townsend, Anthony (2004). "New Digital Geographies: Information, Communication, and Place." In Brunn, Stanley, Cutter, Susan, and Harrington, J.W. (ed.), *Geography and Technology*, pp. 155–178. Dordrecht: Kluwer.

CHAPTER 7

Social Reverberations

This chapter explores several potential issues of adding chatbots to the social fabric by taking the social relationships between humans and chatbots seriously. First, the growing debate about robot rights is being introduced and discussed under the relational approach defended in this project. Second, the principle of human-centered design is being reflected upon to determine whether it is suited to face the challenges arising from human–machine relationships in which the machine may be deserving of some protection in its design process as well. Third, a social split between those who accept and acknowledge human–machine relationships and those who reject these new forms of social interactions is anticipated and outlined in its severity. The fundamental question of how robophobes and robophiles can move forward in incorporating AI agents in everyday life is being explored.

7.1 Robot Rights

The debate about robot rights has been going on for a few years now. There are several accounts to be taken into consideration for a discussion about robot rights, and all of them deserve specific analysis and inquiry. However, considering that in this project, an extensive approach

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_7

has been developed about the reconstruction of human–robot relationships, it seems more worthwhile to approach this debate from its core claims. Additionally, due to the constant influx of technological achievements, the discourse about robot rights is far from settled in forming out "positions" that can be attributed to specific authors. The concentration on four core claims—whether robots can have rights, and whether they ought to—appears to be a safer strategy to cover more ground in this debate.

Thereby, the reconstruction efforts by David Gunkel's "Robot Rights" (Gunkel 2018) may serve as the basis for our elaborations here, with nods toward similar approaches. Gunkel's principal analysis of robot rights proposes to ask any approach two questions: first, whether robots can have rights in the first place, and second whether we should award them rights. Those two questions are distinct, as the answer to the first question does not necessarily determine one's response to the second.

Thereby, this debate generates four possible positions. Gunkel acknowledges that the very discussion about awarding robot rights can be dismissed as a categorical mistake, a position he calls "the unthinkable." (Gunkel 2018, 13) This position cannot be squared with the double rejection of the two guiding analysis questions because it maintains that bringing those concepts together is considered impossible. A fruitful discourse, thereby, is not to be had—it simply does not make sense. Other people rejecting the possibility of reasonably discussing robot rights are Joanna Bryson, who derided the entire debate in an interview as a "waste of time" (Auer-Welsbach 2018) or Noel Sharkey.

A categorical rejection of this magnitude seems inappropriate, as the concepts of "rights" and "robots" are not as firm as this position presupposes. Recent advancements in AI engineering have made the realization of some projections of science fiction more likely, and the discourse about rights has advanced as well. To motivate those rejecting the very debate to participate, it is fair to accept the challenge to lay out the implications of what is meant when saying "rights."

7.1.1 A 4-Way Matrix

In the following, we discuss these four general claims briefly. Because of Gunkel's exhaustive work on the matter, our reconstruction of this debate can remain short. However, all four should at least be mentioned in order

to discuss the underlying issues of the claim of interest for this project to then elaborate on how a relational approach will fit in this debate.

The first position in Gunkel's 4-way matrix rejects the idea that robots can have rights and rejects the notion that they should be given rights. This claim seems to remain within the intuitive notions of rights, as they require specific properties that robots simply will never attain. Additionally, no other technology has even been given serious thought of ever-deserving rights. To defend this position, then, a specific concept of rights has to be justified that may not be shared by others arguing for robot rights.

While opponents of this view may hold a different position on the concept of rights altogether, some may claim that robots will, in fact, be able to attain these properties. This position sometimes argues for creating some sort of servants-class (Peterson 2007, Bryson 2010), in which robots may fulfill all requirements for rights, but ought to be denied those. This position shows a specific interpretation of "rights," as, for most authors, the assertion that something can have rights followed by the fact that these rights are to be recognized while calling for AI agents to be exempt from this normative premise requires a specific concept of "rights." This position has received a lot of pushback, mostly concerning the notion of justifying the institution of slavery by allowing robot slaves. (Estrada 2020) There is, however, a more differentiated position possible that comes closest to the spirit of this position. One could claim that robots may be eligible for certain rights, but the social process of recognizing and awarding those rights should be approached conservatively to avoid misguided projections.

The opposite of a double rejection is a double acceptance. The position here claims that robots should have rights because they can have rights. Most philosophical arguments in favor of this position refer to the expected increase in the abilities of robots over the foreseeable future, leading up to assumptions about an eventual general AI. Defenders of this position thereby can share a similar concept of what "rights" require, as the technological progress will validate those claims. And while nobody at this stage of technological development seriously advocates for immediate recognition of certain protective rights for existing robots, some countries have begun expanding their legal code toward including robots (CIC 2017). Thereby, we can understand this position in the robot rights debate as a preliminary endeavor on how we can approach the inevitable

question of whether highly advanced artificial agents ought to be awarded rights.

Of the highest interest for this project is the remaining position of claiming that even if robots cannot have rights, they still should be considered for being awarded those rights. For this, a different notion of the concept of "rights" is required. If the Kantian premise of normative demands "ought implies can" applies, then claiming that robots should eventually be assigned rights even though they cannot have them is contradictory. That is because assigning rights in the property view includes assigning specific normative capabilities. Rights come with responsibilities, and without the ability to recognize and honor those responsibilities, something should not be awarded rights. This view is a similar point to animal rights. As animals do not recognize any duties due to their lack of human-level speech and consciousness, they cannot be assigned rights. In defending this position against counterarguments, the difference in the use of the term "right" may become evident. Thereby, if we can argue for different uses of the word "rights," then we may reject the verdict that this view is contradictory.

However, first, we ought to examine other arguments brought against this view. In going through Gunkel's collection of arguments against this position, we can find that the relational approach defended in this project is avoiding those. The first argument against this approach to robot rights refers to Bryson's rejection of empathizing with robots (Bryson 2000). The assumption here is that awarding rights to robots is based on a mistaken empathetic move. Similar to empathizing with fictional characters, people can relate to robots, while disregarding the humanity of their fellow humans. Thereby, we require a criterium that is not resting on "moral sentimentalism," but independent from an emotional response.

We can reject this argument on the grounds of our relational approach. As pointed out in Chapter 3, social relationships with robots are not as arbitrary and one-sided like para-social relationships, but interactive and reciprocal. Disregarding the pragmatic relevance of those relationships as purely sentimentalist is thereby a mischaracterization of the motivating factor of the relationships. In fact, the relational, pragmatist approach defended here gives criteria to allow for internal differentiation between human–machine relationships. Thereby, considering assigning rights to those entities is not based on sentimentalism.

The second argument in Gunkel's discussion refers to the anthropomorphic issues with robots. People who are willing to assign rights

to robots may be deceived about the true nature of the technology, as people are projecting human features onto entities that do not have those features. Thereby, awarding robots rights while also holding that they cannot have rights is based on misattributions about the capacity of those robots (Gunkel 2018, 138–139).

We can reject this argument as well, based on the theory laid out before. While it is true that anthropomorphism plays a role in the way people build relationships with robots and animals, it also applies to human–human relationships. We do not call those projections "anthropomorphic," as we know that the features could be found in other people's minds. Instead, we call it "trusting," as we assume that people exhibit certain reliability of their behavior due to states of mind.

Trust and anthropomorphism are the same pragmatic motivation. Both are motivated by the fact that, for the pragmatic reason of guaranteeing our own cooperative efforts, certain assumptions about our cooperation partners are made. And especially considering that those assumptions of trust are not about states of mind (as we have pointed out in Chapter 5, relating to others via mental states is not a working reconstruction), but about cooperation and other pragmatic purposes, the same can be said about certain technological entities. By pointing out that the social relationships between humans and machines ought not to be described as if we were dealing with human–human relationships (Chapter 5), we may even reject the idea that this is an issue of anthropomorphism. To make this point even more explicit, consider the distinction between naïve anthropomorphism and advanced anthropomorphism. Naïve anthropomorphism denotes those projections that are clearly mistaken if taken seriously. Those include worrying about a cleaning robot not getting enough food or that Siri is angry when being approached brazenly. Advanced anthropomorphism is about the ascriptions of certain features based on repeatable, reliable behavior based on a collective, shared understanding (e.g., ascribing a certain "humor" to a dog for enjoying a particular toy). Thereby, these projections are justifiable to a degree since they resemble human behavior to a pragmatic degree. We can find some of these projections in some pet behavior, e.g., there is research into whether dogs can "smile" (Weber 2020), while certainly not all projections of pet behavior are justifiable. When talking about anthropomorphism in legal manners, we usually refer to the latter, and the missing distinction of the grades of anthropomorphism creates a strawman

argument. Advanced anthropomorphism, then, is pragmatically not too different from trusting other humans.

7.1.2 The Relational Turn

This debate about whether robots should be awarded rights while rejecting the idea that they can have motivated some to "think otherwise," as Gunkel puts it. This is not only a pun on the philosophical concept of the Other as previously mentioned, but also a fundamental twist to the robot rights debate as a whole. Coming from Levinas's work on the Other and the relevance of social relationships as a foundation of ethics, Gunkel suggests that the core of considering robots worthy of moral consideration is not their actual, potential, or eventual attainment of moral agency but the quality of relationships they are building with humans (Gunkel 2018, 159f). Thereby, their entrance to the moral circle is not predicated on them being "able to enter," but by the decision of the moral community to add them themselves.

Gunkel explains the definitory issues of determining what a robot is by pointing out that the robots we may want to consider in a robot rights debate are constantly changing and with it our expectations of what robots may become in the future (ibid. 22). Thereby, a debate about "robot" rights is already presumptuous. We noted before that embodiment ought not to play a role in this debate, since our interactions with entities using natural language are the primary phenomenological impact of relevance. The relationships built, with unembodied chatbots or embodied robots, then, do not make a difference in the dimension of properties these entities possess.

Gunkel's proposal, then, is a relational turn toward the Other by reference to Levinas (ibid. 170). Levinas's merit in this debate lies in the uncovering of the method of how Western philosophy was an enterprise largely concentrated on identifying the same, and not the other. Thereby, exclusions from moral consideration were organically flowing from the assumption of who counts as same, and who counts as other. Gunkel moves, with Levinas, to focus on the permissibility of adding entities to moral considerations. The metaphor of having a "face" (Gunkel 2018, 170, Coeckelbergh and Gunkel 2014) is doing some of the work here, as it can be described similar to the phenomenological impact we attributed to NLP agents. In giving robots this "face," a move Gunkel admits is not

provided by Levinas (ibid. 171), they may enter our homes as accepted social members—and thereby deserve moral consideration.

A similar approach has been presented by Mark Coeckelbergh (2010a, 2010b, 2014), who also reconstructs what he calls "moral consideration" of robots by reference to our social relations with them. He underlines the importance of the appearance of certain features as the key part of those social relations. Coeckelbergh states in the well-established phenomenological tradition that we can only claim knowledge of objects as they appear to us (Coeckelbergh 2010b, 214) Thereby, the reliance on the appearance of features ought to be the relevant measure for moral consideration, as the best indirect justification for moral patiency.[1] He acknowledges that those appearances are both context- and subject-dependent, however.

Coeckelbergh moves to include robots in our "social ecology" (ibid. 215) that reach a certain level of relational relevance to us. The term "social ecology" is to be understood similar to the general concept of the "social" or "social fabric" used throughout this book, as it reconstructs the social not only as a horizontal affair of equal agents but as a diverse collection of different forms of social relationships. Adding entities other than humans to this ecology has precedent in both Western as well as Eastern cultural traditions. Through this ecology, we can justify granting moral patiency to certain robots without having to do the laborious and controversial work of defining "robot" beforehand. It thereby allows for the diversity in relational spaces that we have examined in Chapter 5.

These approaches, while providing many similar answers like the one defended here, create some fuzzy edges that ought to be avoided. First, due to the unpredictability of technological developments, any relational approach aiming to establish moral patiency ought to provide a pathway from patiency to agency. That is due to the nature of moral discourse as demanding for discourse participants to take up duties. Eventually, we may want to accept that some robots are capable of being morally obligated, as they are indistinguishable from human moral agents. And while this is not so much a fundamental issue with Coeckelbergh's approach, but rather a theoretical gap, the reference to "appearance" as the hook for relational considerations may not provide the tools to close this gap.

[1] It should be noted that Coeckelbergh does not use the term "moral patiency", but that his efforts to argue for moral considerability of robots through human–robot relationships is very much covered by this term and thereby shall be used here.

In turn, de-centering the human role in moral discourse by referring to specific pragmatic requirements of discourse participants, the "coming of age" of robots from patiency to agency may be incorporated without any issues.

Second, it remains unclear what the role of context- and subject-dependency of those appearances plays in Coeckelbergh's overall approach. In rejecting the idea of a specific approach to evaluating human–robot relationships, an explanation of how exactly an intercultural discourse about recognition of certain relations could look like. Thereby, the requirement of "appearance" seems somewhat of an unwarranted assumption about the relevance of interactive systems. It is entirely plausible to assume that different cultures and sub-cultures will react very differently to the question of which "appearances" are sufficient for moral consideration and which are not. This cultural diversity issue merely reroutes the debate to the subject of moral discourse and who gets to decide which appearances are acceptable.

Third, returning to Gunkel's approach, it seems unclear if the metaphor of "facing" explains the given differences in relating to certain robots. As we argued at the beginning of this book, the ability to talk with us, in opposite to mere behavioral interactions, ought to be considered a fundamental change in creating relational space. The relational habits of humans, learned by the domestication of animals, are not projecting well onto robots. As a result, humans begin relating to speaking robots as if they are relating to human beings. The lack of familiarity with the speaking artificial Other triggers social scripts we only know from relating to the human same, not the animal other. Thereby, the metaphor of "face" and its subtle move toward embodiment and "presence" may be a bit presumptuous here.

7.1.3 The Return of Pragmacentrism

In Chapter 5, Section 5.3. on moral patiency, we introduced a line of thought that may help find a sensible position in the robot rights debate: pragmacentrism. The idea of pragmacentrism was to refer to every moral agent as capable of participating in moral discourse, as those discourses may yield duties and other obligations. Agents, by this understanding, are capable of acknowledging and following those duties, and every entity that is not capable of the latter cannot be considered the former. So far, so

good. With this approach, however, comes the idea of tutoring or chaperoning responsibility of agents toward those entities that are deserving of moral consideration but are not capable of moral agency.

Children and animals have been considered paramount examples of this approach, as they are clearly worthy of moral protection and a place in moral discourse, but not as participants. We alluded to the idea that robots may be added to the list of this group of chaperoned entities by leaning into the relational approach.

Now, how can a pragmacentrist approach navigate the robot rights debate? In establishing moral patiency of robots by burdening moral agents with the task of chaperoning, one can conclude that "rights," understood as the property-based assignments of certain guaranteed treatments, are an unhelpful means to approach the issue. This way, it is in line with Coeckelbergh's and Gunkel's approaches.

The difference between the relational approaches so far and the pragmacentrist idea is that a tutoring approach retains some of the current ideas about agency. However, it connects to Gunkel's and Coeckelbergh's ideas as an extension of how the agency-patiency recognition can be morally motivated in the first place. We have seen that the relationship between humans and robots ought to be understood as a new type of social relationships and we have opened the relational space by comparing it to a second domestication (Darling 2016, 2021 is arguing in a very similar direction). However, in a discourse ethical approach, we require agents to set the rules. Thereby, we do not de-center the moral subject so much as we de-center the conditions of a moral subject and the exclusive rules of being a moral subject.

This way, we can re-approach the debate about robot rights in a somewhat different manner. First, the claim, as put by Gunkel, "Even if robots cannot have rights, they should have rights" can be considered contradictory, as it violates the "ought implies can"-principle. Yet, what is actually asserted in this position is the idea that despite lacking the features required for assigning rights, robots still deserve them as there are other qualifying reasons to assign rights: A normative-legal community is its own authority on whom to grant rights. And while it is understandable that coherence in assigning rights is a valuable feat to avoid randomness in the legal system, there might be outweighing reasons to assign certain rights anyway. Thereby, we can rephrase the claim like this:

> Even if robots cannot have rights like other legal individuals, they still ought to be assigned rights.

However, this still seems unsatisfyingly contradictory. As Gunkel and others point out, the mere discussion about robot rights does not imply that the concept of "rights" is identical to debates about human rights (Gunkel 2018, 51). Extending the legal system to other non-human entities may as well apply different concepts of "rights," e.g., closer to the normative term "claim." If this semantic approach was taken to resolve this contradiction, however, then it seems like another rephrasing ought to be made.

> "Even if robots cannot be the recipients of agent-level rights, they still ought to be assigned patient-level rights."

We can strengthen the first part of this claim again by pointing out that the concept of "rights" does not have to be weakened that much to fit the argument, because we may want to differentiate "the right to be considered" and "the right to act." Darling's point of argument was also not for claiming hard legal rights, but a more mediate approach acknowledging that due to our relations, some robots may be deserving of moral protections (Darling and Hauert 2013, Darling 2016). This recognizes the fact that robots may end up deserving rights after all, as the concept of rights, similar to the concept of "legal status." is not about the transference of human rights to robots, but about the opening of our thinking in legal categories to genuinely new entities. For example, granting robots citizenship seems like a categorical mistake, as citizenship (and its associated rights) is contextualized within a human legal system. For a robot, "national identity" and a guaranteed living environment seem rather pointless concepts (James 2017).

However, some other protections, like the freedom from being destroyed without reason or infected with malware, could reasonably be claimed as a right no robot should be subjected to due to the status of the robot itself.[2] At the same time, it is entirely possible to have those protections in place to protect the relationship of robots with humans.

As the pragmacentrist approach has suggested, there ought to be a pathway from patiency to agency when robots eventually fulfill the

[2] John Danaher calls these obligations "procreative duties.".

pragmatic requirements to join moral discourses—including the acknowledgment of duties emerging from those discourses. Such a pathway would slowly change the justificatory arguments for such protective rights: while currently, the pragmacentrist approach calls for protections to consider robots morally due to their relevance to humans, it may eventually call for similar protections due to the ascribable agency of robots. As a "tutoring approach" to moral consideration, we can expect and encourage any pragmatically capable moral entity to grow from patiency to agency.

It thereby avoids the issues we have encountered with Coeckelbergh's and Gunkel's approach, as a concentration on the discourse pragmatics ought to be agnostic about whether entities other than humans can participate in discourses. Anthropocentrism is not invested, as mere pragmatic requirements for joining a discourse are given.

7.1.4 Conclusion—Robot Relational Protections

The robot rights debate demonstrates the unsettled theoretical landscape of philosophy of technology, especially AI. While some ridicule the very thought of granting robots rights and see a reasonable debate as impossible, others campaign on reworking the legal system now to ensure that robots are incorporated. The crassness of opposites in positions suggests a lack of stable intuitions about commerce with robots will do with us and should do with us. Many of those arguments, then, come from an anticipatory perspective: some fear the reintroduction of a slave class if robots only count as servants. In contrast, others fear the hollowing-out of key properties of the concept of rights if robots do count as legal subjects.

Like Gunkel's and Coeckelbergh's relational approaches, the one defended here does not fit this matrix too well, even though one may count it as agreeing with the sentiment that even if robots cannot be the bearer of rights, they still should count as such. As shown, this statement requires some elaboration to become acceptable. Rights are not necessarily connected to properties since a moral community determines for itself whom to grant rights. And while those properties may eventually be fulfilled, they ought to remain pragmatic. Thereby, a pragmacentrist approach to the robot rights debate can avoid some of the issues encountered in this debate.

7.2 Human-Centered Design

For many debates in interactive systems, the idea of human-centered design and thinking is a guiding principle. The official ISO definition of this principle specifies the approach:

> Human-centered design is an approach to interactive systems development that aims to make systems usable and useful by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, usability knowledge, and techniques (ISO 9241–210:2010(E)).

Essentially, the idea behind human-centered design is the principle that we ought to design any technology with the user of such technology in mind. And especially concerning interactive systems, this approach is applicable to AI and especially chatbot design. This principle allows for the deduction of specific rules of technology development, as well as a shared basis in how to evaluate technological applications. If certain applications are created to exploit its users, then there are clear normative violations.

The human-centered thinking approach is the explicit or implicit core of most moral assessments of technology since most moral assessments are based on human interests. This central position leads to measuring every technology to what it will do to human users and whether humans are also the object of other people's use of said technology.

Additionally, this human-centered thinking suggests a specific interpretation of human–machine relationships, as the design of the latter is always supposed to benefit the former. And considering that all machines are, in fact, designed and made available by humans, those human–machine relationships are ultimately about human–human responsibilities. Creating chatbots under this guiding principle, then, is to commit to creating them in a way that they will not cause harm. We should consider it an open question of whether a human-centered design approach will recommend any sophisticated chatbot as acceptable. Any emotional relationships with chatbots could be disappointed if the suspension of disbelief is interrupted, causing some harm. Those emotional relationships, one could argue from a human-centered design perspective, should be avoided by designing chatbots that do not offer themselves for any relational purpose, i.e., avoiding charming or all too engaging language.

7.2.1 Against Human-Centered Design?

This view, with all the necessary praise for its productive and reliable way of assessing technology, is limited to the categories applied to it. This approach leads philosophers to make a distinction between humans and non-humans that is insurmountable by any other reason but the fact that it is about humans, regardless of the sophistication of the non-human subject. That is, by the very premise of human-centered design, this approach is limiting its moral scope to humans.

The animal rights movement, advocating for awarding legal status to certain highly evolved mammals, serves as a counterweight to this human-based thinking by pointing out a certain relational relevance of animals to the legal system of humans. The argument here is that human legal systems are built on those limiting moral categories to speciesism, like awarding rights to humans merely because they are human, or ascribing abstract concept like personhood based on belonging to a species. The animal rights movement claims, however, that awarding legal status recognizes certain developmental steps, like the capacity to feel pain. Thereby, our legal system ought to extend to the animal kingdom, even if the specific rights that animals should be awarded are up for debate. And if we expand our legal system toward non-human agents, then we clearly ought to consider those non-human agents in designing technology.

This extension of legal categories is not necessarily a clear instruction on how to design interactive systems, as those are merely derived for human use. Thereby, human-centered design is not at odds with animal rights, as their overlap is small. However, transferring the insights of this animal rights movement to technology is not self-evident, as they are both the subject and object of this design. Robots participate in those interactive technologies, and at the same time, are part of those technologies.

As we stated above, it is certainly possible to award robots certain legal protections both based on their relevance for their users (by basically declaring some advanced robots and their software public domain and protection), and also by virtue of their complexity. The concept of "rights" may require some reconsideration, as the term invokes strong intuitions and emotional reactions based on those intuitions, while "protections" or "legal status" may be more productive. Similar to cultural

artifacts like artworks (painting, monuments) and significant performances (like traditions) or natural landmarks (rivers, mountain ranges) (Eckstein et al. 2019), some AI agents may be deserving of legal status.

We may encounter an issue between human-centered design and robot rights, however. Some have begun asking about certain interests of the robot for the robot's sake (like Steve Peterson in the context of sex robots, [Peterson 2017, 155]). If we were to take this position seriously, and we may have to in the future, it would lead to designing technology not only to advance the well-being of human beings but to advance the well-being of everyone interacting with a certain technology. It is unclear how conflicts between the design interests of robots and those of human beings can be resolved, even though it is already foreseeable that these conflicts are going to be fought with emotional verve.

This conflict does not seem to materialize in the way normative conflicts usually materialize. In making some design choices for humans, the product may be a robot that is "suffering" in the sense that its interests are being ignored, possibly in the form of a sex robot. However, anticipating the robot's interests before creating it presumes that those created interests cannot *not* be created. If we equip a robot with the ability to feel or suffer, and we put it to use in tasks that will cause this suffering, we might as well argue that the equipment of said suffering is misplaced. Measuring this suffering against the human suffering by decentering design from humans, the design choices of robot construction are taken into account as if they were an equal factum with the ability of humans to suffer. That is simply not the case.

The motivations for creating more and more capable robots to the point where they might start plausibly claiming certain capabilities that ought to be recognized is to serve human purposes. If we construct a robot that is surpassing human purposes by competing with those purposes, its development simply has overshot the needs of technological development (similarly argues Thomas Metzinger (2009) that we should not create a robot that can suffer). Thereby, human-centered design within the robot rights debate can provide a theoretical limit as to how far robot development ought to go: the moment robot needs and human needs are measured against each other, we ought to reject the robot's needs. However, the social-relational, pragmacentrist approach suggests that there are reasons to take robot needs into consideration. That is the case when some human needs are depending on those robot needs.

7.2.2 Robot Relations, Human Realities

There is another side to human-centered design that may interfere with the high-flying goals of robot construction. The human factor in creating and applying artificial intelligent agents is often overlooked as a condition for the success of technological promises. This is true for some PR-strategies (like the case of what was presented as an advanced humanoid robot ultimately just being a man in a robot suit, Roth 2018) as well as the everyday use of algorithms to improve the efficiency of certain work processes, often to the detriment of human workers (Valovic 2018).

As some argue (e.g., Birhane and van Dijk 2020), the categories of debating AI ethics ought not to be AI agents vs. humans, but corporations vs. workers. Presuming that no robot will have interests to be taken seriously, they urge to understand the economic ramifications of the current developments. Robots are not the innocent inventions of humble engineers in their garages, but the product of giant corporations with enormous power to exploit given economic structures to their advantage. Without taking into consideration the human toll of developing artificial agents, the debate about robot interests is already biased: one could conclude, in fact, that human-centered design is already failing those who are supposed to implement the new possibilities of AI. From overworked delivery drivers to permanently surveilled office workers, without casting a wide enough net about the effects of AI on society, the debate about robot rights is leaving out crucial parts of human society.

The relational approach taken here is not directly contributing to this debate, as it concentrates on a specific type of AI, the communicative, and those human agents that are open and willing to enter relationships with them. However, this does not mean that only one of these debates is valid; both concerns, the one about the economic circumstances of AI development and the one about the relational space of human–machine interactions, are normatively legitimate as they both pertain to humans interests.

The former indeed affects the other, as has been analyzed in Chapter 4, Section 4.5. However, ignoring the progress made on the front of relational spaces and emerging human–machine relationships will not help anyone and only will lead to further alienation of those who do see meaning in relating to artificial beings. On that ground, both areas of thought have their place in the contemporary philosophy of AI.

7.3 Including Non-Human Agents in Normative Systems

The move of proposing to include non-human agents is a logical consequence of accepting that chatbots will not be meaningfully different in their skill, but also a consequence of opening the possibility of relating in a genuinely new way, as has been elaborated in the previous chapters by redefining the social as a relational network.

If chatbots gain skills to talk with us to the degree that allows for a sufficiently big group of people to relate to those chatbots in a sufficiently convincing manner (that is, convincing for those who do not relate to those chatbots but can understand the merits of such a relation) or if chatbots gain those skills so that people relate to those chatbots in a manner that is on par in relating to pet animals, and if those relationships are determined to be fundamentally different from any human–human relationship and any human-animal relationship, then a new category is required to outline the circumstances, conditions, and consequences of those new human–machine relationships. This argument is the building block to establish the human–machine social category. Some of the interactions between robots and humans have been pointed out already, like the need to understand the term "friendship" between humans and machines as a possibly unworkable category not in practice but in the essentialist presumptions and the history of the term as exclusively denoting certain human–human relationships.

The relational space opened with this idea is not empty. However, it also still requires some positive answers as to what those new relationships within this social category could look like.

7.3.1 Relational Real Estate

As we elaborated before, some philosophical approaches to understanding human–machine relationships as neither anthropomorphic nor dismissible are somewhat vague about the extent and quality of these social categories. It does not seem like robots will reach sufficient levels of all human capabilities any time soon. However, it would seem like a mistake to rest on the assumption that a simple "never going to be human" would suffice in rejecting any social convention-building effort to include robots. Both the fact that people will be starting to relate to technology socially, as well

as the philosophical support for some of those relations, renders a dismissive "not-human argument" toothless. We may all be very well aware that those relations are not at all like human–human-relationships, but an insistence on that may not suffice to convince those in relationships with machines to value them differently. Whether we agree with this development or not, we better prepare for a future with active human–machine relationships.

7.3.2 Legal and Digital Persons

In the following, one proposal of how we can socially categorize chatbots and other artificial agents in our social fabric is put forward. The main goal for this chapter is that we can work on developing the relational space with new descriptors to populate the human–machine relational landscape.

The suggestion made here might help with recognizing a certain status of those robots that is neither entirely unprecedented nor coming from a given social circumstance: the digital person. It is common in legal theory to recognize "legal persons," which are usually institutions like corporations or nations, to be considered the subject of a certain action, contract, or otherwise. "Legal persons" are thereby culturally tested entities representing certain collective actions. Through collective action, in which no individual action can be identified as the "main" or "core" agent, questions of responsibility are diffused and limited. The discussion about the ontological status aside, there is a certain plausibility to treating "legal persons" as if they were actual acting individuals in the context of responsibility and legal ramifications of collective actions. However, whether individuals or collectives are the adequate addressees of blame for legal or moral transgressions depends on strict institutionalized rules.

A "legal person" is thereby a social construct which, for many purposes, we are happy to ascribe agency and liability to, if not even degrees of responsibility (see also Gellers 2020 for a legal theory on the transfer of rights to non-human entities). And while this only extends to the normative realm of legal theory and practice, similar concepts have been discussed in the area of technological development. One can follow several approaches in detaching the issue of agency and responsibility from singular, atomic agents (like humans), e.g., Verbeek's "joint responsibility" (Veerbek 2009) and Nissenbaum's "many hands" approach (Nissenbaum 1996). With these, not only legal persons can be

considered a workable entity in certain normative contexts, but digital persons in certain social contexts.

While with legal persons, there is no debate on whether they should have rights similar to natural persons,[3] they are still acknowledged as "addressable" in a communicative sense. People in Germany tend to complain about the German train corporation without having any person specifically in mind—they are referring to the legal entity of the corporation. Thereby, it seems like legal persons can be, to a certain degree, considered similar to natural persons. However, ontologically there are clear differences between legal and natural persons. That is to show that for introducing the term "digital person," the ontological status does not have to be cleared.

We could even go further by pointing out that "digital persons" are, in general, closely related to legal persons in their blame-structure. If a chatbot insults someone, like the Twitterbot Tay has done (for an analysis of Tay's case, see Liu (2017), we can to a degree blame the Twitterbot (without the urgent feature of having a blame subject, as the functional theory of blame seems to suggest by referring to the functional role of blame of directing responsibility, not as a state of mind that requires a blame subject, see McKenna [2012]), but also the engineers that constructed and published the bot. The idea of joint responsibility of autonomous technology and its creators allows for a transitional blame-structure in which technology starts sharing some of the blame without the blame being misdirected.

Legal persons, as the established example of an amalgamation of collective action, do not always behave in a way that would be predictable for any involved agent; similarly, digital persons will behave in ways that no engineer could reasonably predict. Yet, with legal persons, we are comfortable assigning blame. Thereby, it seems also reasonable to expect that we can assign blame to digital persons as both entities in their own right as well as a subject to blame a group of collective agents.

[3] An exception might be the current precedent set in the "Citizens United v Federal Election Commission" case of 2010, in which the Supreme Court of the United States of America ruled in favor of the group "Citizens United" to grant legal persons the right to free speech, and counting donations as speech, concluding that corporations spending money on political actions is covered under the first Amendment to the US constitution. Similarly stipulates the German Constitution ("basic law", Grundgesetz) that all basic freedoms also apply to legal persons "as far as the nature of such rights permits" (Basic Law 1949, §19:3).

Ontologically speaking, legal persons are a different category to natural persons, but still very much "real" in the sense that they are part of our shared reality and have effects on us as much as other institutions and organizations. And if the ontological concept of a legal person is not causing philosophical stomachache, then it seems hard to argue how digital persons, as similar composite entities, should be treated any differently.

7.4 New Social Fault Lines

A relational approach, like the one presented here, is not limited to human–machine relationships. It might appear that way, as establishing the relational space for human–machine relationships is the primary goal both for this approach as well as some others (Gunkel 2012, Coeckelbergh 2010b, 2014, Darling 2017) However, there is an entirely different dimension to this debate that ought to be reflected upon here—and that is its consequences on human–human relationships.

The question of how the presence of chatbots and robots in society will affect human–human relationships has two sides: on the one side stands the issue of those who reject any form of human–machine relationships. In an increasingly digitized world, more people than ever before concluded that the constant use of digital media and participation in this second-layer reality is exhausting, overdemanding, and ultimately not in their best interest. The speed with which technological innovation is changing people's lives has caused some to reject those innovations altogether, e.g., the movement of "Neo-Luddites" who advocate for activism to slow down technological innovation (Jones 2006).

On the other side stand people who are open and willing to adopt new technology. For our relational approach, the most interesting are those agents who are establishing relationships with chatbots and perceive them as meaningful. Eventually, they ask to be taken seriously in those relationships (or, at least, in their perception of them). There have been cases of para-social relationships that made headlines, as people claimed to be in romantic relationships with the Eiffel Tower (Simpson 2008) or other objects. The move of pathologizing those para-social relationships as mental disorders will reach its limits when the depth of chatbot interactions is not reasonably described anymore as a "subject-object relationship."

Yet, these two dimensions are not distinct, and some innovative technological applications may be reaching people who otherwise reject technological progress. For example, care robots or self-driving cars could hope for high acceptance rates with people who otherwise would depend on other humans to help them with their tasks. The potential for increased mobility and care for people in an age group usually risk-averse might change their minds on some concrete applications, not on the technological infrastructure as a whole. They might be using a self-driving car to see their grandchildren but reject any form of innovation in their respective areas of interest.

Thereby, the new fault lines will not emerge between humans and robots, but between humans themselves. To elaborate on those fault lines and how to propose to ameliorate them, we require some distinctions.

7.4.1 Rogers' Bell Curve and Its Limits

Currently, sociological distinctions between technology usages and interactions are usually classified among Rogers' bell curve first published in "Diffusion of Innovation" (Rogers 1962), in which Rogers differentiates between innovators, early adopters, early majority, late majority, and laggards. This characterization has helped classify different approaches to new technology by focusing on the usage and involvement of users with a given innovation (usually a certain type of technology). Innovators and early adopters are signifying a small, technology-forward minority and the majority of technology users being late adopters ("late majority" and "laggards" are constituting roughly 50% of technology users in Rogers' estimate). This approach has been a very popular and successful basis for explaining and predicting certain uses of technology among a population. It also has been elaborated to include social circumstances of an individual's attitude toward certain technology, like access to said technology, age, general technophobic or technophile social surroundings. With these inclusions, we can construct more accurate models about an individual's preference for using or not using certain technologies at certain stages.

However, with certain types of attitudes toward and use of technology come social ramifications for any user (or non-user). As pointed out above, many issues between humans rejecting new technology can be found here. The term "laggard" is often understood from the side of early adopters as a term of derision and is cause for generational divides. For example, the use of the app "Snapchat" has been characterized as a

marker for a generational gap (Alvarez 2015). In part, due to its unique approach to communication, even some early "digital natives" do not connect too well with this type of technology. This cut-off creates a social ecology of mostly young media users under the age of 25.

And while generational gaps highly correlate with the acceptance rates of new technologies (ibid.), this does not apply to all technology users, as some applications. Especially in medical technology, many otherwise laggards are keen early adopters (Ketikidis et al. 2012).

Turning to chatbots and their anticipated development, we can notice these gaps as a common occurrence: media users growing up with chatbots to talk to will receive them fundamentally different from older users. Their perception of chatbots as part of the "objective spirit," i.e., the culturally given reality, will result in fewer questions and philosophical issues as people may have right now. It is highly doubtful whether people growing up with a more elaborate personal assistant will question its ontology and whether the relationship they have with it may be fake or insincere.

Thereby, those who do not participate in interactions with chatbots might become alienated from those who are comfortable interacting with communicative AI. However, abstinence from the use of certain technology and the resulting split in society may be explainable with the common denominations of Rogers' bell curve: the difference in approach technology, whether favorably and open or averse and unfavorably, has been an issue before communicative AI has emerged, and thereby is not a genuinely new issue.

However, the opposite side of this development will not be explainable this way and creates a genuinely new social problem. According to Rogers, innovators and early adopters usually make up a minority of technology users. In the characterization from above, those innovators and early adopters ought to be understood as developing a new kind of relationship with technology that goes beyond mere "adoption": they will establish human–machine relationships perceived as meaningful. Due to natural-language processing allowing chatbots to keep intellectually and possibly emotionally fulfilling conversations, these relationships will reach social relevance.

So far, the relational approach to human–machine relationships can explain how those relationships could be understood and socially categorized. However, the gap between laggards and innovators, i.e., the social distance between those who enter social relationships with robots

and those who reject robots as relatable altogether, will encounter a hard conflict.

The more complex and convincing chatbots become, the less persuasive the insistence becomes that robots are no entities we can relate to. It becomes an issue of whether we should relate to them in the ways some people practice. Additionally, every instance of improved skills of chatbots will make a pathologization of human–machine relationships less plausible, as has been pointed out above as well. This conflict is thereby becoming a normative one, as the conflict is primarily not about people's attitudes toward robots and the idea of human–machine relationships, but about people's attitudes toward other people's attitudes to human–machine relationships.

7.4.2 On Robophobes and Robophiles

The attitudes of people toward speaking robots are not limited to their own use of technology, but to other people's use of said technology as well, requires a new distinction between those who reject these kinds of human–machine relationships and those who are willing to accept them.

As this conflict will potentially emerge and linger, as many social conflicts about human–human affairs are centuries old, the framing of this conflict is important. It is explicitly not an issue between humans and robots. For a long while, robots will not be capable of making any reasonable assertion about themselves unless we want them and make them to. We have seen this issue in the debate around robot rights and human-centered design: creating robots that demand rights for their own sake already presumes that we have left the human-centered design approach that we ought not to leave.

It is an issue between humans, instead. Or, more precisely, an issue between attitudes. The naming convention available for these conflicts in attitude suggests the suffixes -phile and -phobe, as they reflect the generally positive or generally negative attitude people can have. In connection with the robot rights debate, the subject may best be robots. Thereby, we can call the one side "robophobes" and the opposite side "robophiles."

Robophiles are willing to invest more emotional and cognitive capital in their interactions with robots or are accepting of those who do. Robophobes deny the possibility of building meaningful relationships and thereby are committed to considering robophiles as being mistaken about their investments. If someone denies the fact that humans can

have meaningful relationships with robots, such as friendships or merely trusted acquaintances, then they are committed to dismissing anyone who claims to have established such a relationship and thereby what I call "robophobic."

Robophobia, then, could be the philosophical consequence of having a strong theory of mind. However, this does not require robophobes to be aggressively anti-robot, as the term may imply. Rather, robophobes will not recognize any human–machine relationship as meaningful in a stronger sense of the word other than one of a service provider (robot) to a recipient of services (human). There are plenty of examples in the current literature, most prominently Noel Sharkey (Sharkey 2018), even though the term "robophobe" has not been used to describe them.

Robophiles, however, will claim to be able to establish such meaningful relationships, like friendships, or at least accept and support anyone who does. The currently unclear definitions of what a "meaningful in the stronger sense" human–robot relationship can look like, both in a professional context as well as in the general population, contribute to misconceptions that may fuel resentments and prejudice. It appears that most people's threshold for a meaningful human–machine relationship is informed by science fiction, which usually portrays robots with unscientific assumptions about theories of mind and our computational abilities.

However, looking closer into the psychological ramifications of what it means to have a "meaningful" relationship, we might want to add that reciprocity is not necessarily limited to intentional give-and-take. For example, "respecting someone" and "being respected by someone" are entirely definable within the perceptive limits of the affected agent. If certain behavioral protocols are performed, an agent might judge that they are being respected, even though those behavioral protocols are performed disinterestedly. In this case, then, a robot may "respect" agents the same way humans can, without assuming certain states of mind behind the respecting entity.

Obviously, the two positions of robophobia and robophilia are not just absolute but allow for inner differentiation. One can reasonably have high requirements for accepting human–machine-relationships as meaningful by expecting them to be nearly on the level of reciprocity as human–human-relationships. These requirements would put someone on the robophobic side—not in absolute terms, but relative ones. Similarly, robophiles can differ in their interpretation of meaningful human–robot

relationships. While some may be very quick to accept the attribute "meaningful" if someone claims their relationship is meaningful, possibly bordering on para-social relationships, others may put up some intersubjective conditions of reciprocity, durability, or, as presented in Chapter 3, Section 3.1, embeddedness.

7.4.3 One Battleground: Robot Rights

The question of whether robots can and should be awarded rights represents a good test for the introduced distinction between robophobes and robophiles, as it is both a hotly contested academic question, as well as a matter for possible future civil rights movements. As stated at the beginning of this chapter, many areas of the robot rights debate only extend to human–machine relationships, not to robots themselves.

The difference between the extreme positions of the groups within the robot rights debate lies partly within the assumed possible relational space. While those discussions are currently limited to academic evaluations and conceptual debates about the best use of the concept of rights, some of the issues mentioned and discussed above are going to enter the public debate rather soon. The lack of alternatives between robophilic attitudes and robophobic ones may cause severe disruptions in society, as one side will eventually move forward, making political demands for which the other barely sees any reason to pursue.

Gunkel's point in bringing up robot rights in the first place is not to advocate for them in any specific way, but to point out that ascribing and recognizing other people's rights is a relational action between humans. The fact that working-class people, women, people of color, homosexuals, and other groups have been denied civil and political rights for most of the path of Western civilization is not grounded in the fact that those groups did not have any rights, but that they were denied rights by not recognizing them. The question emerging from this approach is whether or not we will eventually come to a point where recognizing robot rights will be a hard question to face or not.

The relational approach defended here suggests a similar approach to Gunkel, even though it recognizes that legal protections of certain tutoring relationships are appropriate. Considering that the law ought to protect certain freedoms and the well-being of the citizens living under the rule of law, protecting relationships perceived as worth protecting is a prime task of the legal system.

Due to the latest developments within legal frameworks to grant robots certain rights (CIC 2017), the debate about robot rights will soon face the first hard tests of the current legal systems. Both sides have begun activist groups to shape the debate in their goals, with groups advocating for the ban of certain types of robots (like for sex robots (Campaign Against Sex Robots (2015) or killer robots (Stop Killer Robots [2012]), while others take a supportive role of introducing robot rights (ASPCR 1999).

7.4.4 A New Struggle for Recognition

While a relational approach to robot rights may help to establish some moral terminology, it still is an open question of how this terminology can be effectively placed in a discourse about the acceptability of human–machine relationships and the norms we can build on those relationships.

One way of looking at this discourse is by introducing a theory of justice that can help find norms of acceptability to guide future debates about human–machine interactions and their recognition in society.

For this purpose, the theory of recognition by social philosopher Axel Honneth (Honneth 1992) is an adequate entrance due to its focus on recognition of identities and relationships within the bigger context of power structures. Honneth claims that for the constitution of society, the social recognition of the other is a constitutive part. Recognizing others as what they are, while taking into account the circumstances of their identity construction, is, according to Honneth, a constitutive part of a just society.

In applied ethics, this theory has been considered a complementary element to other theories of justice, which have been concentrating on mere distributive and legal elements, like Rawls' Theory of Justice (Rawls 1971) or some legal positivism. The theory of recognition requires moral agents to acknowledge other people's contextual and complex subjectivity, and not seeing them as a compound of individual preferences and desires, like other approaches in social theory and theories of justice have done.

Human subjects, to put it concisely, are more than the sum of their preferential parts: they constitute complex contextual identities that are not encapsulated by distributive justice. Thereby, only creating institutions that provide an individual with all their reasonable needs while

guaranteeing the overall freedom of society, like Rawls' idea of a just society presumes.

The focus on the complexity of human subjects puts expressed identities at the center of the recognition-approach to justice. This does not work without investigating the conditions of the expression of identities, i.e., its power structure, and social conditions (Honneth 2003). To create a just society from the perspective of recognition theory, then, it is paramount to allow for the expression of identities, be it cultural, sexual, gender, religious, or otherwise, and organize institutions in a way that does not suppress individual identity expressions, but support them.

Recognition theory has its roots in critical theory, as critical theorists like Lukacs (Lukacs 1967) have reflected upon the conditions of self-relations and self-expression in capitalist societies. Lukacs's main concern is the reification of the self and of relationships between people; that is, the perception of one's self as a thing due to the commodification of the relationships with others. Lukacs, unsurprisingly, identifies the capitalist social structure as the source for commodification of the relationships with others and the reification of the self.

These sources allow for understanding power structures, economic conditions, social circumstances, and individual self-expression as intimately connected—if the economic conditions are leading individuals to perceive themselves as things, a human connection based on recognizing the other as a human being is impossible.

With the importance of recognition as a social mechanism contributing to justice in place, we can assume that recognizing relationships that are considered meaningful by those who are participating in it constitutes an element of justice toward robophiles. A robophobic position, then, seems problematic on the grounds of recognizing a human being's subjectivity as experiencing a human–machine relationship as real as other relationships.

It would be just to call for recognizing those relationships as meaningful and worthy of some institutional protection, similar to the protection of pets not only as a property of their owner but also as an entity with significant emotional value to the owner.

However, we could argue on the robophobic side that not the recognition of human–machine relationships is the issue, but the reification begins with human–machine relationships. Considering that machines are things built with a capitalist logic in an effort to make money, building

relationships with them must be transactional and contributes to the reification of the self as Lukacs's fears. Thereby, the reification of the self is amplified through relating to chatbots that are not individuals but are personalized machines produced for a market.[4]

This argument makes some assumptions that we rejected earlier. First, it supposes an essentialism about social relationships that we argued against in Chapter 3, Section 3.3. Whether or not a relationship with robots is perceived as merely transactional or not is not for anyone to judge but the person in the relationship. We may warn them about the conditions and limits of a human–machine relationship, but those limits are not an essential qualitative marker of relationships, but the ones we are used to and deemed worth pursuing.

Second, the issue of human–machine relationships is not solved by pointing out the reification of the self and the capitalist conditions under which the reification unfolds. Nobody is forced to enter those relationships, but we may expect from anyone who does not that recognize the human's independent identity of having relationships with robots. Denying others this recognition based on one's own assumption of the reified self in this relationship lead to similar paternalistic and robophobic arguments we have ruled out above.

7.4.5 Limits of Recognition

Yet, the theory of recognition also encounters some unique limits in the case of human–robot relationships.

One argument could be that nobody is born with an inclination toward human–machine relationships. Thereby, those who engage with machines on a level that demands recognition from others do so while rejecting the humans around. Choosing one side can always be interpreted as rejecting another side, and nobody can be expected to grant more than basic courtesy to those rejecting them.

However, we can reject this argument on two grounds. First, we simply do not know whether future generations will perceive robots in such a natural way that some develop natural inclinations for them over human companionship. Those "natural robophiles," then, are not choosing their preferences but perceive them as given. Second, rejection does not forfeit

[4] This argument is owed to Sebastian Nähr-Wagener.

one's claim to recognition. A hermit living all on their own devices may still demand recognition of their choices and identity. In fact, those who are rejecting us are merely the tough cases where a recognition-approach comes into effect: recognizing those in their identity, autonomy, and self-expressions that are close to us is easy. Recognizing those that are not is more difficult.

Another argument is that the recognition of those who relate to robots requires acknowledging the subtle power imbalance between humans and machines. Speaking machines are, in the end, able to be much more than just natural-language processing companions; their design and ability allow them to record every conversation, extract data, like user patterns and preferences, and share those with the parent company. The fact that in the current situation of AI design and production, a chatbot's purpose is to increase user engagement, provides an insight into the severe imbalance of power that should inform the relatability of chatbots. Siri or Alexa, even when personalized to fit a specific user's preferences, will share some sort of data with its parent company, and is not bound to the conversational customs we built based on human conversational limitations (a similar point make Newman & Blanchard 2019).

Thereby, any user developing a relationship with robots will most likely also be part of a one-sided power relationship with a strong corporation. In the same perspective of being cautious about the constructing purpose of chatbots, we should demand from human beings to develop and maintain a certain autonomy and level-headedness, something Shannon Vallor calls "technomoral wisdom" (Vallor 2016). This wisdom-mediated autonomy should guarantee a person's control and agency in every human–machine relationship. The choice of those entering human–machine relationships cannot be preserved by going down a robophobic route of insisting that their relationships will never be meaningful, but by gently reminding them that those relationships are fundamentally different in nature than established human–human relationships. If someone accepts that fact, we can argue, we ought to recognize those relationships and should seek regulation to protect them as well as we protect human-pet relationships.

Thereby, the recognition of those human–machine relationships should not be unconditional, as this is a slippery slope toward para-social entities that might ultimately be harmful to the person in it, too. We can expect from humans willing to engage with machines a certain awareness of entering a relationship with an entity principally different from us.

The individual responsibility for exerting Vallor's "technomoral wisdom" when engaging in technology of any kind, is also called for when engaging in these new kinds of relationships.

7.4.6 Some Ethical Deliberations

As human–machine relationships become more complicated, intricate, and harder to distinguish from human–human relationships, the stakes are rising in how to find a social consensus about acknowledging these relationships. Their uniquely unfamiliar quality might pose some problems in evaluating the available options of how to deal with them.

However, there is a normative distinction to be made between robophobia and robophilia that is worth spelling out. The normative concerns robophobes may have are based on two related issues: the first, somewhat paternalistic concern is that people might be getting hurt because they relate to machines in a way that is not (and may never be) covered by the machine's complexity. Human–machine friendships require involved reciprocity and intention, and many philosophical theories suggest that we are far away from implementing those in robots. Being mistaken about one's friendship with a robot can hurt due to the projection of features of human–human friendships, and this hurt is increased if we promote the idea that the transfer of friendship features is possible in the first place.

Second, robophobes might be concerned that a liberal stance toward human–machine relationships eventually forces them to interact with robots in a way they reject. This consequentialist argument assumes that if human–machine relationships are recognized as such, robots will find their way into the social fabric of almost everybody. The base expectation of how to interact with robots may change to a robophobe's disadvantage. If human–machine relationships are accepted as possibly meaningful relationships, and Roger's bell curve moves forward, then the laggards are in a situation where their refusal to recognize and build relationships requires justification due to the "new normal" established by widespread acceptance. Even if robophobes gave up their first concern, that robophiles will be hurt by committing to misattributed relationships, then this second will remain: they will be required to in some form accept human–machine relationships, even if they refuse to engage with robots themselves.

Arguments against the first concern are rather striking. A robophobe's concern for someone else's vulnerability requires a strong paternalism that

seems hard to justify. If we take other humans in their autonomy seriously and assume that they have developed the "technomoral wisdom" Vallor demands, then their choice to enter those human–machine relationships and the risk associated with these relationships ought to be respected. The argument to control other people's choices to enter relationships is hard to justify in the first place, and actively interfering with someone's wishes to relate to non-human entities requires pathologizing of those relationships. With every instance of increased features of chatbots and robots, this pathologizing move can be rejected, and with it, the paternalistic argument.

The second concern appears more substantial. As robots become more and more ubiquitous in everyday interactions, equipped to maintain and reaffirm certain social customs, they will become part of the social fabric. Similar to seeing someone with their dog on the street, we will encounter people talking to their chatbot friends or coworkers. Robophobes will be forced to acknowledge those robots as social entities, even if they believe they should not be. And while it seems unlikely that there will be an argument to force robophobes to give up their attitudes toward robots, the issue lies with the recognition of other people's relationships with robots. However, as long as nobody is forced to establish social relationships with robots themselves, the argument only extends personal attitudes. We have discussed the issue of being mistaken about human or chatbot conversations in Chapter 5, Section 5.1. as a downside of the paradigm of anthropomorphism. Nobody should be mistaken about whether they are talking to a human or a machine. But with those rules in place, being against the changing overall attitudes toward a possibly more robophilic society is not a sufficient argument.

7.4.7 A Positive Argument

Here, a more general point can be made. This debate mirrors some experiences with movements demanding recognition, from women's suffrage to the Black civil rights movement to the homosexual and queer liberation. The demand of those movements has been (and arguably still is) aimed at gaining recognition of their existence, their rights, their identity, and their relationships. And while a robophilia-driven movement may seem to be about demanding the recognition of intrinsic worth and rights to robots, it is actually about the relationships of humans with those robots. With the theory laid out here, we do not have to refer to robots as

subjects and bearers of rights but can think of them as important factors in people's social lives.

If the historic struggle of recognizing other people's relationships has taught us anything, then it is better to recognize relationships too early rather than too late. Homosexual or interracial relationships long have been dismissed as somehow wrong, often based on the same dynamic: personal attitudes toward the same sex or other races are being transferred to judge and dismiss other people's relationships. The safer strategy to defuse potential future conflict between robophobes and robophiles is to take this lesson from these historical struggles and err on the side of magnanimity and acceptance.

7.5 Conclusion

The divide between those who accept and use the next stage of natural-language processing AI and those who oppose such technology is bigger than mere technology adoption. The fundamental changes not only to society as a human–human enterprise but as a richer network of agents that are on the horizon are out of the scope of Rogers' bell curve. Thereby, new terminology ought to be introduced to delineate the divide between those who believe that meaningful human–robot relationships are possible and those who do not.

This paper proposes to distinguish between robophobic and robophilic attitudes. The idea behind this distinction is not to focus on the reservation people have toward robots, as those reservations depend on highly contingent design choices and engineering paradigms, but pointing out that this divide is whether some people's claimed relationships ought to be recognized or not.

For this, it is not helpful to point at the extreme cases of people developing para-social relationships. In Chapter 3, Section 3.9, the difference between social relationships and para-social ones has been clarified. Whether a human–machine relationship is being recognized as a social relationship or dismissed as not a full relationship, is not only a philosophical question but will be a political one. Similar to the robot rights debate and the unfolding political movements there, the issues about accepting others and their choices to rely on machines for their social relationships will be a concern for public debate.

Recognizing those relationships may contribute to justice, as often enough history has shown an early recognition of relationships is the position to prevail in the long run. There are some valid concerns on the paternalistic side, as the power imbalance between humans and machines is built on the difference in how chatbots "remember" events and "know" certain facts that between human–human relationships would not occur. Thereby, the creation of sociable machines ought to be guided. However, this does not suffice as an argument that justifies fundamental opposition to human–machine relationships, which is presented by some robophobic authors as the only "reasonable" way to react.

References

Alvarez, Hannah (2015). The Generation of Snapchat. UX for Different Age Groups. *Usertesting*. https://www.usertesting.com/blog/snapchat/ (Last accessed on August 27, 2024).

ASPCR (1999). *American Society for the Prevention of Cruelty against Robots*. http://www.aspcr.com/index.html.

Auer-Welsbach, Christoph (2018). 15 Minutes with Leading #AI Specialist Joanna Bryson. *Medium*. https://medium.com/cityai/fifteen-minutes-with-leading-ai-specialist-joanna-bryson-c944b7c3fd25 (Last accessed on August 27, 2024).

Basic Law of the Federal Republic of Germany (1949). Translated by: Professor Christian Tomuschat, Professor David P. Currie, Professor Donald P. Kommers and Raymond Kerr, in *Cooperation with the Language Service of the German Bundestag*. https://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html#p0105 (Last accessed on August 27, 2024).

Birhane, Abeba and Van Dijk, Jelle (2020): Robot Rights? Let's Talk about Human Welfare Instead. In *Proceedings of the AIES Conference 2020*.

Bryson, Joanna (2000): "A Proposal for the Humanoid Agent-Builder's League (HAL)." In Barnden, John and Lee, Mark (eds.), *Proceedings of the AISB 2000 Symposium on Artificial Intelligence, Ethics and (Quasi-) Human Rights*, pp. 1–6. https://aisb.org.uk/wp-content/uploads/2019/12/AISB00_Ethics.pdf (Last accessed on August 27, 2024).

Bryson, Joanna (2010). "Robots Should be Slaves." In Wilks, Yorick (ed.), *Close Engagements with Artificial Companions*, pp. 63–74. Amsterdam: John Benjamins Publishing Company.

Campaign Against Sex Robots (2015). https://campaignagainstsexrobots.org/ (Last accessed on August 27, 2024).

Campaign to Stop Killer Robots (2012). https://www.stopkillerrobots.org/ (Last accessed on August 27, 2024).

CIC (2017). *Saudi Arabia is First Country in the World to Grant a Robot Citizenship*. Press Release, October 26, https://cic.org.sa/2017/10/saudi-arabia-is-first-country-in-the-world-to-grant-a-robot-citizenship/ (Last accessed on August 27, 2024).

Coeckelbergh, Mark (2010a). Moral Appearances: Emotions, Robots, and Human Morality. *Ethics and Information Technology* 12(3), 235-241.

Coeckelbergh, Mark (2010b). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology* 12(3), 209-221.

Coeckelberg, Mark (2014). "Robotic Appearance and Forms of Life: A Phenomenological-Hermeneutical Approach to the Relation Between Robotics and Culture." In Funk, Michael and Irrgang, Bernhard (eds.), *Robotics in Germany and Japan: Philosophical and Technical Perspectives*. Frankfurt a.M.: Peter Lang.

Coeckelbergh, Mark, and Gunkel, David (2014). Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. *Journal of Agricultural and Environmental Ethics* 27(5), 715-733.

Darling, Kate (2016). "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects." In Calo, Ryan, Froomkin, A. Michael, and Kerr, Ian (eds.), *Robot Law*, pp. 213–234. Cheltenham: Edward Elgar.

Darling, Kate (2017). "'Who's Johnny?' Anthropological Framing in Human-Robot Interaction, Integration, and Policy." In Patrick Lin, Keith Abney, and Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, pp. 173–192. Oxford: Oxford University Press.

Darling, Kate (2021). *The New Breed. How to Think About Robots*. Penguin.

Darling, Kate, and Hauert, Sabine (2013). Giving Rights to Robots. *RobotsPodcast #125*. https://robohub.org/robots-giving-rights-to-robots/ (Last accessed on August 27, 2024).

Eckstein, Gabriel, D'Andrea, Ariella, Marshall, Virginia, O'Donnell, Erin, Talbot-Jones, Julia, Curran, Deborah, and O'Bryan, Katie (2019). Conferring Legal Personality on the World's Rivers: A Brief Intellectual Assessment. *Water International* 44(6-7), 804-829. https://doi.org/10.1080/02508060.2019.1631558.

Estrada, Daniel (2020). *Tweet from January 21, 2020*. https://publish.twitter.com/?url=https://twitter.com/eripsa/status/1219672216935321601 (Last accessed on August 27, 2024).

Gellers, J. (2020). *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. Routledge.

Gunkel, David (2012). *The Machine Question: Critical Perspective on AI, Robots, and Ethics*. Cambridge, MA: The MIT Press.

Gunkel, David (2018). *Robot Rights*. Cambridge, MA: The MIT Press.

Honneth, Axel (1992). *The Struggle for Recognition*. Suhrkamp: Frankfurt am Main.

Honneth, Axel (2003). "Redistribution as Recognition: A Response to Nancy Fraser." In Fraser, Nancy and Honneth, Axel (ed.), *Redistribution or Recognition? A Political-Philosophical Exchange*, pp. 110–197. New York: Verso.

James, Vincent (2017). Pretending to give a Robot Citizenship Helps No One. *The Verge*. https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia (Last accessed on August 27, 2024).

Jones, Steve E. (2006). *Against Technology: From the Luddites to Neo-Luddism*. Boca-Raton (FL): CRC Press.

Ketikidis, Panayiotis, Dimitrovski, Tomi, Lazuras, Lambros, and Bath, Peter. (2012). Acceptance of Health Information Technology in Health Professionals: An Application of the Revised Technology Acceptance Model. *Health Informatics Journal*, 18, 124-34. https://doi.org/10.1177/1460458211435425.

Liu, Yuxi (2017): The Accountability of AI—Case Study: Microsoft's Tay Experiment. *Chatbotslife*. https://chatbotslife.com/the-accountability-of-ai-case-study-microsofts-tay-experiment-ad577015181f (Last accessed on August 27, 2024).

Lukacs, Georg (1967). *History and Class Consciousness*. London: Merlin Press

McKenna, Michael (2012): *Conversation and Responsibility*. New York: Oxford University Press.

Metzinger, Thomas (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York (NY): Basic.

Newman, Daniel, and Blanchard, Olivier (2019). *Human/Machine. The Future of Our Partnership with Machines*. London: Kogan Page.

Nissenbaum, Helen (1996). Accountability in a Computerized Society. *Science and Engineering Ethics* 2, 25-42.

Peterson, Steve (2007). The Ethics of Robot Servitude. *Journal of Experimental and Theoretical Artificial Intelligence* 19(1), 43-54.

Peterson, Steve (2017). "Is it good for them too? Ethical Concerns for the Sexbot." In Danaher, John, and McArthur, Neil (eds.), *Robot Sex: Social and Ethical Implications*, pp. 155-172. Cambridge, MA: The MIT Press.

Rawls, J. (1971): A Theory of Justice. Harvard University Press: Cambridge (MA).

Rogers, Everett (1962). *Diffusion of Innovation*. Simon and Schuster: New York.

Sharkey, N. (2018), "Mama Mia It's Sophia: A Show Robot or Dangerous Platform to Mislead?", Forbes, https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#407e37877ac9 (Last accessed on August 27th 2024).

Simpson, Aislinn (2008). Woman with Object Fetish Marries Eiffel Tower. *Daily Telegraph*. https://www.telegraph.co.uk/news/newstopics/howabo

utthat/2074301/Woman-with-objects-fetish-marries-Eiffel-Tower.html (Last accessed on August 27, 2024).

Vallor, Shannon (2016). *Technology and the Virtues. A Philosophical Guide to A Future Worth Wanting*. Oxford University Press: Oxford.

Valovic, Tom (2018). Workplace AI: A Dystopian Nightmare in the Making? *The Sociable*. https://sociable.co/technology/workplace-ai-dystopian-nightmare/ (Last accessed on August 27, 2024).

Veerbek, Peter-Paul (2009). "Cultivating Humanity: Toward a Non-Humanist Ethics of Technology." In Berg Olsen, Jan-Kyrre, Selinger, Evan, and Riis, Søren (eds.), *New Waves in Philosophy of Technology*, pp. 241–263. Hampshire: Palgrave Macmillan.

Weber, Lisa Beth (2020). Do Dogs Smile? The Science Behind the Looks We Get From a Happy Dog. *PetMD*. https://www.petmd.com/dog/behavior/do-dogs-smile-science-behind-looks-we-get-happy-dog (Last accessed on August 27 2024).

CHAPTER 8

Large Language Models as Relational Technology

LLMs, general-purpose chatbots, and any more specific interactive conversational technology will be more socially relatable than any other technology before. This has thus been the premise under which philosophers have speculated about questions of moral standing, robot rights, and other terms to reckon with their different social and moral status. David Gunkel has been one of the leaders in this regard, rejecting ontological arguments that categorize robots (or, in our case, conversational machines), with reference to the lack of clarity in the choice of the relevant concepts, the content of the relevant concepts, and the relevance of the relevant concepts (Gunkel 2023). This has led to a debate on whether ontological status is discernible and, if so, whether it should count, for moral standing considerations, especially when it comes to sentience (e.g., Dung 2024; Jaworska and Tannenbaum 2021).

As we pursue a relational account here, the relevance of ontological features of technology can be relegated to other areas of philosophical debate. However, in taking such a perspective (and implicitly rejecting others), we ought to be justified: the relational account of moral standing, which considers robots (and conversational machines) relevant moral patients by extension of their relationship to us (and ultimately upending the subject-object relationality of moral standing) and rethinks our relationship to technology altogether, can remain neutral on the outcomes of

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_8

these debates while avoiding to be incurring the moral risk of incurring harm before the debate is settled.

8.1 The Issues with "Do No Harm"

LLMs can say all sorts of things, and one challenge for both AI companies as well as the public at large is to determine behavioral rules and limits for what kind of utterances chatbots ought to be allowed to make, and what kind of safety limits we must establish to keep them from uttering problematic statements.

"Avoiding harm" as the sole moral concern for the utterances of a chatbot falls short of reaching the ethical permissibility of a chatbot but also considerably harms its relatability. The concept of "safety" is highly reductive when applied as the sole normative framework for chatbots. In this context, safety is often defined as the avoidance of anything that could potentially harm the users or the wider social fabric. However, this account is overly simplistic and does not account for the complexities of social and moral norms, which often require more than just avoiding harm. For instance, in some situations, it might be safe for a chatbot to avoid taking a stance on a controversial issue, but such neutrality might not always be the most morally or socially responsible course of action. In fact, a more engaged response—one that is careful and aware of the controversy—might be required to align with broader moral and social expectations.

Moreover, this reductionist view of safety leads to difficulties in assessing the relative moral and social quality of different "safe" responses. Two answers that are both deemed safe may still have vastly different implications, which might be significant to the user. If a chatbot is judged solely on safety, these nuances are overlooked, and the final assessment is left to the subjective interpretation of engineers or crowd-workers, who may not fully grasp the complexities of the situations in which the chatbot operates.

8.2 Appropriateness

As a consequence, other approaches have been developed that complement the "do no harm"-rule from AI safety considerations, including the concept of "appropriateness" (Kasirzadeh and Gabriel 2023; Kempt et al. 2024). For our purposes here, the conceptual offer from Kempt,

Lavie and Nagel will work best, as they present three different forms of appropriateness relevant for discussions surrounding chatbot-utterances: technical-discoursive appropriateness, social appropriateness, and moral appropriateness. With these distinctions in mind, we are able to capture the moral guardrails (and sometimes the specific moral content) that chatbots ought to adhere to or provide.

8.2.1 Technical-Discursive Appropriateness

Technical appropriateness captures the condition of what usually is considered discourse ethics. As a technology able to participate in discourses (and having elaborated upon the premise of pragmacentrism for discourse ethics in Chapter 6), the adherence to some of the basic rules for rational discourses is a condition applying here. This means a highly sophisticated uptake and detection of even finer texts and contexts of human speech (and speech acts) is required: being able to properly assess the state of conversation is a prerequisite for any social machinery (cf. Kempt 2022), and the ability to aptly and appropriately respond to a diverse, quickly changing, and at times inconsistent conversational protocol is a skill LLMs still lack.

Being able to appropriately respond to conversational states, however, also implies the ability to appropriately remain consistent with oneself. As a source of frustration (or ridicule), current responses from LLMs and general-purpose chatbots appear to prioritize the pleasing of the user, rather than being discursively appropriate. The ability to stay consistent with oneself, and be sensitive to evidence and true statements, is a tacit assumption we make about people joining discourses. This is evidenced by the common reaction of giving up trying to reason with someone who seems to be insensitive to true statements, or who is constantly contradicting themselves in an attempt to make a larger, ideological point. These rules are, in the conceptualization of discourse ethics, a rational necessity for any purpose-driven, shared discourse. General-purpose chatbots, however, have yet to demonstrate their ability to be consistent with their previous statements (or even within one utterance).

8.2.2 Social Appropriateness

Social appropriateness, in turn, is a larger set of norms, usually associated with context-specific rules of behavior within a subgroup. Similar to

technical-discursive appropriateness, social appropriateness is usually not about content but about behaviors, but not about the rational pragmatics of discourse, but about the social contexts in which one behaves. Standard examples include manners and patterns of speech, which vary culturally and are usually considered relative to the context in which one speaks. Some of these norms are created within moral appropriateness; however, many cases (e.g., when to thank whom and to what degree for a favor that was done) are in fact limited to habits and commerces that are a result of cultural, local, or otherwise un-generalizable. This includes difficult to foresee developments in subgroups, terms that change, slurs that get reappropriated within subgroups, new norms of interactions emerging in generational cohorts, etc.

8.2.3 Moral Appropriateness

Moral appropriateness addresses the ethical obligations we owe to each other in a conversation. While safety covers some moral norms, such as avoiding harm, it does not account for positive moral duties, such as providing comfort or support when needed. For example, if a user expresses grief over a personal loss, a chatbot programmed only to avoid harm might disengage from the conversation to prevent generating unsafe outputs. However, this would be morally inadequate, as we would expect a human interlocutor to offer some form of empathy or support. Moral appropriateness, therefore, requires the chatbot to go beyond merely avoiding harm and to actively contribute to the well-being of the user. Further, moral appropriateness is a more robust approach to avoid socially ambiguous, potentially problematic statements. A general-purpose chatbot, as suggested here, is not a member of a subgroup but may represent its own subgroup. Thus, social appropriateness requirements might fail to take hold for chatbots, while moral appropriateness-norms, as well as technical ones, apply to every discourse participant.

With these concepts in place, we ought to discuss some of the drawbacks we might encounter here. These include the standardizing of social norms, the highly context-dependent nature of appropriateness, and the resulting fictionality of some contexts that are open to lying, curse-words, and other elements that would be considered “inappropriate.”

8.2.3.1 The Moral Risks of Standardizing Social Norms

One of the challenges in implementing appropriateness as a normative framework for chatbots is the risk of standardizing social norms across diverse cultures. This issue arises particularly in the context of "corporate mainstreaming," where chatbots are trained based on the norms considered acceptable in specific corporate environments, such as those prevalent in the United States. This approach can lead to the imposition of certain cultural standards on a global scale, potentially marginalizing other cultural norms and values.

For example, if a chatbot is trained to follow the norms of politeness as defined by corporate America, it might not perform well in cultures where different norms of politeness apply. This could result in misunderstandings or even offense in cross-cultural interactions. Moreover, the anthropomorphization of chatbots—where users begin to treat them as human-like agents—can exacerbate these issues, as users may expect chatbots to adhere to the same cultural and social norms as human beings.

8.2.3.2 The Context-Dependence of Appropriateness

Another important aspect of appropriateness is its context-dependence. The norms that govern appropriate behavior can vary significantly depending on the setting and the purpose of the conversation. For instance, certain statements that might be considered inappropriate in a formal conversation could be entirely acceptable in a fictional or educational context. This variability poses a challenge for designing chatbots that can navigate different conversational contexts without violating normative standards.

Consider the example of fictional storytelling, where it might be appropriate for a chatbot to use graphic or disturbing imagery to create a compelling narrative. In this context, the chatbot's behavior is governed by different norms than it would be in a real-life conversation. However, this flexibility also opens the door to potential misuse, as users could exploit these context-dependent norms to bypass safety measures and prompt the chatbot to generate inappropriate content.

8.2.3.3 The Challenge of Counterfactuals and Lying

The issue of context-dependence becomes particularly problematic when dealing with counterfactuals and hypothetical scenarios. For example, should a chatbot be allowed to lie in a fictional context, such as creating a story where one character deceives another? While lying is generally

considered morally inappropriate in real-life interactions, it might be necessary for creating realistic fictional narratives. This raises the question of how to balance the need for creative flexibility with the ethical requirements of truthfulness and honesty in chatbot interactions.

Moreover, the ability of chatbots to engage in counterfactual reasoning—imagining scenarios that are not true or have not yet occurred—can lead to complex ethical dilemmas. For example, a chatbot might be asked to generate a scenario in which using offensive language is justified by an extreme situation. While some might argue that this is inappropriate, others could see it as a legitimate exploration of moral reasoning. These controversies highlight the need for careful consideration of the ethical implications of chatbot behavior, particularly in situations where the boundaries between reality and fiction are blurred.

8.3 Positionality: Chatbots *as* *Chatbots*

As we have seen in the previous chapter, the lack of social positioning can be identified as a reason for chatbots behaving more like humans than chatbots. It appears that the artificiality of the conversational AI only emerges at the limits of what it is programmed to do—either by a technological limit or by a normative one. This is because chatbots are not anchored in a social setting in which they have a default role they can revert to.

Before we can discuss the difficulties and requirements to find a social space of chatbots as chatbots, we should elaborate on the concept of "positionality," especially in contrast with other, currently deployed terminology of placing chatbots in discourse. Positionality refers to the specific and recognized *subjectivity* of a conversation participant. In our context, it means the normative requirement for a chatbot to maintain an established discursive and social position that is also recognized by those interacting with the chatbot. This concept is related to Google's "role-consistency" metric, which assesses how consistently chatbots maintain an assumed role (e.g., speaking from a specific perspective about objects), and the metric of "groundedness," which supposedly grounds the utterances of chatbots in the appropriate contextual assumptions (as we have seen before, the performance of most general-purpose chatbots is lacking in this metric due to hallucination issues).

However, while role-consistency focuses on maintaining a role, positionality defines the "default role" a chatbot should have, which then can

be manipulated in a shared context. Google's engineers may prioritize consistency in perspective, factual accuracy, and not breaking character. In contrast, we argue that because "positionality" involves ensuring the social and moral appropriateness of any given discursive position, it is essential to clearly define the default position a chatbot should hold. This could mean equipping general-purpose chatbots like ChatGPT, Gemini or personal assistants like Alexa and Siri with a specific point of view or discursive role while restricting them from assuming certain other roles. The prohibition of taking other roles is key, as those roles will be performed according to consistency requirements, possibly reverting to harmful utterances.

Currently, this is partially addressed by chatbots declining to provide opinionated answers on sensitive topics, given their nature as Large Language Models (LLMs). However, since a chatbot can take on various roles upon request, it must be subject to social and moral appropriateness for those roles and be limited in adopting certain roles. This approach requires positionality without requiring the chatbot to take a definitive stand. Whether the sophistication of language models should influence their moral consideration is an entirely different issue. Positionality simply requires us to metaphorically "carve out a space in social contexts" to establish norms and expectations for chatbots' appropriate roles.

Research on the moral relevance of chatbot roles, due to their perceived relatability (e.g., Gunkel 2022; Coecklebergh 2012; Danaher 2019a), highlights the importance of avoiding anthropomorphization and other potentially misleading or harmful projections. Chatbots should not adopt typical human traits: see for example the discussion surrounding the permissibility of chatbots using emojis to convey the meaning of some of their utterances (Véliz 2023). This is a clear example of problematic anthropomorphism, in which some conversational cues humans use to relate to each other have simply been incorporated into artificial speech, seemingly giving users a reason to project human traits onto the entity.

Instead, they should have a position within discourse without taking stances on content-related questions or occupying roles typically held by humans. This is distinct from the debate on whether chatbots should be considered as friends (Danaher 2019b; Kempt 2022) or lovers (Nyholm and Frank 2019). While the question of the relatability of a chatbot *as chatbot* is relevant for questions previously only extending to human relationships, positionality is foundational to these debates.

Psychological research indicates the empirical salience of these theoretical debates. As people accept robots in discursive settings and appear to feel excluded if a robot is not paying "attention" to them (Rosenthal-von der Pütten and Bock 2023), it seems that the perception of robots (or other conversational machines) is being integrated into our conversational norms even if robots are not anthropomorphized—their position *as machines* achieves the same.

The guiding insight is that chatbots are not intended to replace human interlocutors by mimicking or pretending to be increasingly human, but rather to expand conversational norms and spaces to include chatbots *as chatbots*. To ensure appropriateness, we need to reach a consensus on how to interact with these machines, what roles they should have, and what kinds of relationships we want to establish with them. Providing chatbots with a default positionality and point of view will help connect them to the concept of appropriateness.

8.4 Other Values and Requirements

Appropriateness is a useful concept to capture the output we ought to demand from an artificial speaker that may just say anything. However, in order for it to integrate into larger ethical discourse, we require more detailed ideas about what our relationship with this technology should consist in, and how the question of value alignment ought to be addressed. For this, we first turn to the idea of "acceptability" as a minimum requirement for users to interact with the technology, and second to the question of value alignment (cf. this discussion in Kempt et al. 2023).

8.4.1 Acceptability

To better assess what role chatbots should play in society, we should also consider the concept of "acceptability," as we want to enable the adoption of such technology into a wider social setting. Technology assessments and the normative conclusions derived from them have relied on rationalist principles. These principles define an "acceptable" outcome as one that can reasonably be expected to be accepted by all users of the technology, including the inherent risks associated with using this technology. Factors such as the individual and societal risks associated with a

technology, its benefits, and its widespread use all contribute to this assessment. The more ubiquitous a technology becomes, the more essential it is to manage its risks effectively. Acceptability, therefore, can help determine the appropriate stance a chatbot should take, guiding us in evaluating the acceptability of the remaining risks associated with such technologies.

Moreover, we ought to acknowledge that certain user requests may be deemed unacceptable. Just as we would reject inappropriate requests in a face-to-face conversation, we should expect chatbots to reject requests that are based on unacceptable input. This distinction highlights the inadequacy of using "safety" as the sole criterion for evaluating chatbot interactions, but also the lack of "bite" of "inappropriateness" as a moral (or social or technical) category of the chatbot. If we only focus on the chatbot's output, only responses that are potentially unsafe would be flagged as problematic. However, if we take the chatbot's future role in conversations seriously and require it to have a discursive positionality, we must also consider the morality of the requests made to the chatbot. In other words, asking a chatbot to do something unacceptable should be part of the moral evaluation process.

This means that we ought to consider both the permissibility of the output of a chatbot (i.e., its safety and appropriateness) and the acceptability of the user input and any appropriate response to such input. Acceptability, being informed in part by social research on acceptance of some technology or its features, also helps to reduce the foreseeable tension between those willing to socially incorporate chatbots, and those taking a robophobic stance (see also Chapter 6, Section 6.4).

8.4.2 Value Alignment

The issue of Value Alignment, which has been a major concern in long-term discussions about AI (Gabriel 2020; Hristova, Magee & Soldatic 2024; Kirk-Giannini 2024), especially in the context of Large Language Models, is closely related to positionality, acceptability, and appropriateness. The alignment problem is one of the most significant challenges in AI development: how do we ensure that AI applications align with their intended purposes (which we assume to be aligned with our general values), and how do we prevent AI from constantly misaligning, especially in the realm of reinforcement learning-fine-tuned LLMs that appear too fragile to guarantee reliable alignment (Shen et al. 2023)? This is relevant as the AI, from a certain level of sophistication onward, is subject

to constant evolution that often lies outside of human hands (see, for example, the time chatGPT became "lazy," as the AI learned that in its internal reward-system incomplete answers are preferable to wrong ones [Mahdawi 2024]). To maintain an "aligned AI," we must first agree on the purposes that these AI applications are supposed to serve. This question of alignment is central to the broader discussion about the role of AI in society.

However, this insight brings us back to a concern we sought to avoid: currently, it is not that AI is misaligned with "our" purposes, but rather with the purposes of corporations. Yet, these purposes—whether short-term, mid-term, or long-term—often conflict with the needs and expectations of a public that is overwhelmed and disoriented by rapid technological changes. Without a clear use case for general-purpose chatbots, the worry arises that corporate goals will drive the alignment process, rather than the public interest. This could lead to a situation where chatbots are aligned more closely with corporate interests than with what is genuinely "best for the public." Therefore, the alignment question may be more relevant when investigating the alignment of AI companies themselves, rather than focusing solely on AI alignment. This perspective is supported by recent analyses that emphasize the importance of considering the broader political economy of LLMs and the role of corporate goals in shaping AI development (Luitse and Denkena 2021).

It is important, however, to acknowledge that constructing chatbots for large-scale use will likely require some concessions from the public—regardless of who builds them. Similar to how the introduction of automobiles necessitated a renegotiation of public spaces, the introduction of chatbots into everyday conversations will have downstream effects that require adjustments. Parking spaces, highways, traffic lights, traffic deaths, speed limits, etc. all were necessitated by a new technology that simply was too useful to not integrate into our everyday life. Generative AI—or rather some applications that are based on such AI—might mark a similar change in our attitudes toward the public discourse. Positionality, which involves assigning specific roles and expectations to conversational AI, represents one such adjustment.

This re-alignment process will require a broader public debate about the roles and responsibilities of chatbots in society. For chatbots to be socially and morally appropriate, it is essential to consider the diverse cultural and linguistic contexts in which they operate. Chatbots must be able to navigate different cultural norms and social expectations,

which may vary significantly across regions and communities. Without this sensitivity, chatbots risk perpetuating biases and reinforcing harmful stereotypes, leading to negative consequences for users.

8.4.3 Context-Dependency and Context Variance

As laid out above, in addition to social appropriateness, moral appropriateness is a critical consideration for chatbots. Moral norms go beyond simply avoiding harm; they encompass positive duties that conversational agents must fulfill. For example, if a user expresses sadness or grief, a morally appropriate response would involve offering support or providing resources, rather than simply avoiding potentially unsafe outputs. Moral appropriateness requires chatbots to demonstrate empathy and understanding in their interactions, reflecting the ethical obligations that human interlocutors would have in similar situations.

However, the concept of appropriateness is highly context-dependent. What may be appropriate in one conversational context may be inappropriate in another. For instance, fictional or educational settings often allow for speech that would be considered unacceptable in real-life conversations. This context-dependency presents challenges for general-purpose chatbots, which must be flexible enough to adapt to different conversational scenarios without compromising on appropriateness.

One of the key concerns in this regard is the potential for chatbots to lie or engage in counterfactual reasoning. While it may be morally inappropriate for a chatbot to lie to a user in response to a direct question, there are situations where lying or creating fictional scenarios may be necessary or even desirable. For example, chatbots may need to generate fictional stories or engage in role-playing exercises that involve deceptive elements. The creation of LLMs that can engage in counterfactual reasoning, or at least behave technically appropriately to counterfactuals is a key to navigating contexts (Li et al. 2023). However, this flexibility also opens the door to potential misuse, such as generating harmful or misleading content. Therefore, it is crucial to establish clear guidelines and boundaries for when and how chatbots can engage in such behaviors.

The concept of appropriateness, encompassing social, moral, and technical-discursive dimensions, offers a more comprehensive framework for evaluating chatbot behavior, even in deliberately fictional contexts. By considering acceptability, value alignment, and the contextual nature of appropriateness, we can better navigate the challenges of integrating

chatbots into society in a way that respects both individual and collective values.

8.5 Conclusion

The fact that LLMs appear to be a promising base for the creation of increasingly more relational technologies, i.e., those that can maintain psychologically stimulating, socially integrateable, and normatively appropriate conversations, poses new challenges for the philosophical work needed to be done beforehand. Next to several other relational accounts, we focused on the concepts of appropriateness and positionality to provide orientation.

The distinction of three different kinds of appropriateness helps to identify the different oddities in utterances of a chatbot as a chatbot, even when they are not harmful. Technical appropriateness merely states that chatbots must follow conversational rules, while social appropriateness stresses the culturally relative sentiments and norms about how to properly behave. Lastly, some behavior is relevant on a moral level, e.g., the appropriate reactions to someone else expressing certain feelings. This last one we called "moral appropriateness."

Positionality, then, cements the requirement to be aware of the potentially problematic issues as laid out in Chapter 5, Section 5.1 on the political and moral implications of anthropomorphism. Imitating human beings is burdened with normative issues, and finding a unique position from which chatbots may speak is a key to embed them into wider social conversations.

References

Coeckelbergh M. (2012). *Growing Moral Relations*. London: Palgrave Macmillan. https://doi.org/10.1057/9781137025968

Danaher, J. (2019a). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviorism. *Science and Engineering Ethics,* 1–27. Online first at: https://link.springer.com/article/https://doi.org/10.1007/s11948-019-00119-x

Danaher, J. (2019b). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies* 3(1), 5-24

Dung, L. (2024). Preserving the Normative Significance of Sentience. *Journal of Consciousness Studies,* 31(1), https://doi.org/10.53765/20512201.31.1.008

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds & Machines* 30, 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gunkel, D. (2022) The relational turn: thinking robots otherwise. In: Loh J, Loh W (eds) Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds with Robots. Bielefeld: Transcript, 55–76.

Gunkel, D. (2023). Person, Thing, Robot. A Moral and Legal Ontology for the 21st Century. MIT press

Hristova, T., Magee, L., and Soldatic, K. (2024). The Problem of Alignment. *AI & Soc*. https://doi.org/10.1007/s00146-024-02039-2

Jaworska, A., and Tannenbaum, J. (2021). The Grounds of Moral Status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/. (Last accessed on August 27, 2024).

Kasirzadeh, A., and Gabriel, I. (2023) In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philos. Technol.* 36, 27. https://doi.org/10.1007/s13347-023-00606-x

Kempt, H., Lavie, A., and Nagel, SK (2023). Appropriateness is All You Need! *arXiv*. https://arxiv.org/abs/2304.14553

Kempt, H., Lavie, A., and Nagel, SK (2024). Towards a Conversational Ethics of Large Language Models. *American Philosophical Quarterly* 61(4).

Kempt, H. (2022). *Synthetic Friends. A Philosophy of Human-Machine Friendship*. Palgrave.

Kirk-Giannini, C.D. (2024) Is Alignment Unsafe? *Philos. Technol.* 37, 110. https://doi.org/10.1007/s13347-024-00800-5

Li, Y., Xu, M., Miao, X., Zhou, S., and Qian, T. (2023). Prompting Large Language Models for Counterfactual Generation: An Empirical Study. *arXiv preprint* arXiv:2305.14791.

Luitse, D., and Denkena, W. (2021). The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI. *Big Data & Society*, 8(2). https://doi.org/10.1177/20539517211047734

Mahdawi, A. (2024). What is Going on with ChatGPT? *The Guardian*. https://www.theguardian.com/commentisfree/2024/jan/12/chatgpt-problems-lazy (last accessed August 27, 2024).

Nyholm, S., & Frank, L. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to "love" their owners? Techné, 23(3), 402–424. https://doi.org/10.5840/techne2019122110

Rosenthal-von der Pütten, A. & Bock, N. (2023) Seriously what did one robot say to the other? Being left out from communication by robots causes feelings of social exclusion. *Human-Machine Communication*, 6117(134). https://doi.org/10.30658/hmc.6.7

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., ... and Xiong, D. (2023). Large Language Model Alignment: A Survey. *arXiv preprint* arXiv: 2309.15025.

Véliz, C. (2023). Chatbots Shouldn'T Use Emojis. *Nature*, *615*, 375

CHAPTER 9

Conclusions

9.1 Thinking Forward

Before we take a look at what lies ahead, we should look back at the lines of thought we followed to get where we are right now.

In the chapter about social relationships, we explored the relevance of ascribing social descriptors, putting social relationships into a pragmatic context. Instead of assuming predetermined social categories and concepts like an essentialism does, the view proposed here reconstructs social relationships as the flexible characterizations of socially recognized interactions. These interactions are conditioned on technological progress and availability.

Domestication and digitization were identified as being two ways technology has changed our understanding and thereby our characterization of social relationships. Domestication made it possible to enter social relationships with animals, the interpret their relationship with us based on a genuinely new social category. Digitization, in turn, showed that not only new types of social relationships are possible, but that the essentialist definition of some common relationships, like friendships, is inadequate.

The chapter on conversational artificial agents was intended to be less philosophical and more a reflection on the technological progress of artificial intelligence. The specification of dealing with chatbots specifically made such a reflection on the intelligence necessary. Without

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3_9

a solid understanding of the technological processes and progressions involved in the production of chatbots, the relationship of natural-language processing with other forms of AI, and the current method of teaching algorithms how to speak, the philosophical consequences would have been untethered from reality.

The concentration of possibly unembodied conversational agents is motivated by the conviction that language represents the constitutive part of the human mind, and that sophisticated communicative AI will have a fundamentally different impact on our perception of AI as "an Other" than humanoid robots will.

We saw that language models are growing in size and sophistication. Due to practically unlimited available language data from often unaware internet users, these models could be considered "general" in the sense that chatbots will soon be fully open-domain, i.e., convincing conversation partners.

With both a concept of social relationships and speaking machines at hand, an exploration of social relationships with those speaking machines was possible. The core issue of chatbots is the norm of anthropomorphism in their construction. The imitation game of human speakers has led many to react skeptical toward their use, as issues of deception, embarrassment, and imbalances of power can occur. To ameliorate those issues, we proposed to create chatbots that are distinctly non-human. This non-anthropomorphism comes at the cost of disorientation. If human speakers are not the goal of imitation, how would we design speaking machines?

Some of the common philosophical approaches were discussed, especially the relationship of those chatbots to Hegel's idea of an "objective spirit." With the introduction of the points of domestication and digitization, we found that chatbots may have a similar effect to lead to genuinely new, yet substantial and meaningful relationships as domestication did. Domestication has not put forward one singular category of relatable animals. In fact, it has brought a variety of different relationships, of which only a few are social relationships. This variety of relating to animals can be transferred to AI agents. Just because something is artificially intelligent does not tell us much about its relatability. However, language does, and thereby chatbots. This lays the groundwork for approaching human–machine relationships between humans and speaking robots from a supportive perspective: many human–human relationships lack physical presence thanks to the digital possibilities of connecting with people all around the world. Communicative skills like being a good listener, the

ability to keep secrets, and availability for one's needs are of high importance for many people and could be implemented in chatbots without requiring anthropomorphistic strategies.

The recognition of human–machine relationships left us with a challenge many philosophers have tackled: the debate on whether robots should be awarded rights. However, through the line of thought walked down in this book, this debate appears somewhat odd. While many discourse participants argue whether robots could ever be agents to be deserving of those rights, or if they will ever have mental states of such complexity that they cannot be ignored any longer, our relational approach had some straight-forward answers to the initial question. As have David Gunkel, Mark Coeckelbergh, and others claimed before, this debate ought to be about the relations built between humans and robots, not about robots as a decontextualized subject. In fact, this decontextualization does not make much sense from a relational perspective, as without those relationships, this debate would lack motivation. The fact that "robot rights" is an issue in the first place is that there are some people credibly claiming to be entering social relationships with robots. The relational approaches in this debate focus on those relations, not on the robots themselves. This book proposed to understand these relationships as one of tutoring. The recourse to discourse ethical assumptions provided us with a tool of de-centering an agency-centered approach to ethics and established patiency to those entities that otherwise would not be taking part in moral discourse: pragmacentrism.

Pragmacentrism focuses on the pragmatic abilities of those participating in a discourse, and until robots fulfill these (or if they never), we still can take care of them due to their relevance to other agents. Representing robots in discourse does not require embodiment of any kind, thereby reaffirming the relevance of the relational part over the specific robot.

However, this position may also come with some challenges if human-centered design is put into consideration. The construction of a robot with needs and skills to suffer may be inherently immoral if we cannot guarantee that it will not constantly suffer; however, those robots would certainly create more intimate relationships with humans. Eventually, this development may lead to weighing a robot's interest against a human's interest. We argued here that such a weighing process ought to be undercut. Taking a relational approach seriously commits us to insist on the relationship as the basis for moral consideration of the robot.

Lastly, we reflected on how human–machine relationships will affect human–human relationships. Not only will the ubiquity of artificial intelligent technology lead to people rejecting the technology in many forms altogether. However, this strong rejection and those new forms of relationships people form with a technology entirely rejected by others will certainly create tension between advocates and critics of chatbots.

Since these are substantial issues about the extent to which human–machine relationships ought to be protected, an adequate naming convention would be one of robophobes and robophiles. It reflects previous struggles for recognition of those entering human–machine relationships as a genuinely new form of social relationships. Moreover, as a recommendation of how we should treat these issues of intimacy and identity, we found that erring on the liberal side of those issues has been a prudent move to avoid harming vulnerable populations.

9.2 Acting Forward

The initial reaction of adults toward Aibo and their concern that children playing with a plastic toy would lead them to be desensitized toward animals was mistaken. The children did not lose their category of social relationships to animals due to an inanimate object. Instead, they seemingly effortlessly expanded their understanding of social relationships toward this barking piece of plastic. This goes to show that our social categories are open and expandable if we allow them to be. The lack of certain skills or assumed mental states will not keep people from building trusted relationships with chatbots, or use them as their go-to listeners to vent about work or heartache. Chatbots are used to help autistic kids find a conversation partner with the endless patience a human being would not have (Newman 2016) or are used to translate complicated text to easy-read or other simplified language to make the world and its information more accessible (Freyer et al. 2024). The options for using talking machines in our midst are endless once we break away from anthropomorphism as a guiding rail and tool of rejection for the construction and acceptance of chatbots.

The technology is there. The arguments are there. The theory is there. Whether the social acceptance is there, remains to be seen. As one consequence, the philosophical work cannot stop in finding a strong argument in the robot rights debate but communicate it to those who are skeptical of the entire process. Psychologists may pathologize humans entering

human–machine relationships similar to the pathologization of excessive internet use without understanding that the use was an outlet for vulnerable populations to connect with similarly affected people. Legal scholars may have a particularly rigid understanding of what "rights" are and reject any notion of protecting an artifact because of its significance to the artifact's owner. Similar to the animal rights debate, the robot rights debate demands some action to change some minds. A relational approach may provide the argumentative tools to do just that. In a nod to David Gunkel's approach to "think otherwise," we may want to encourage to "think forward."

References

Freyer, N., Kempt, H., and Klöser, L. (2024). Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. *Ethics Inf Technol* 26, 50. https://doi.org/10.1007/s10676-024-09792-4

Newman, Julia (2016). *To Siri, With Love*. London: Quercus.

Index

H. Kempt, *Chatbots and the Domestication of AI*, Social and Cultural Studies of Robots and AI,
https://doi.org/10.1007/978-3-031-75558-3