How to build nice robots

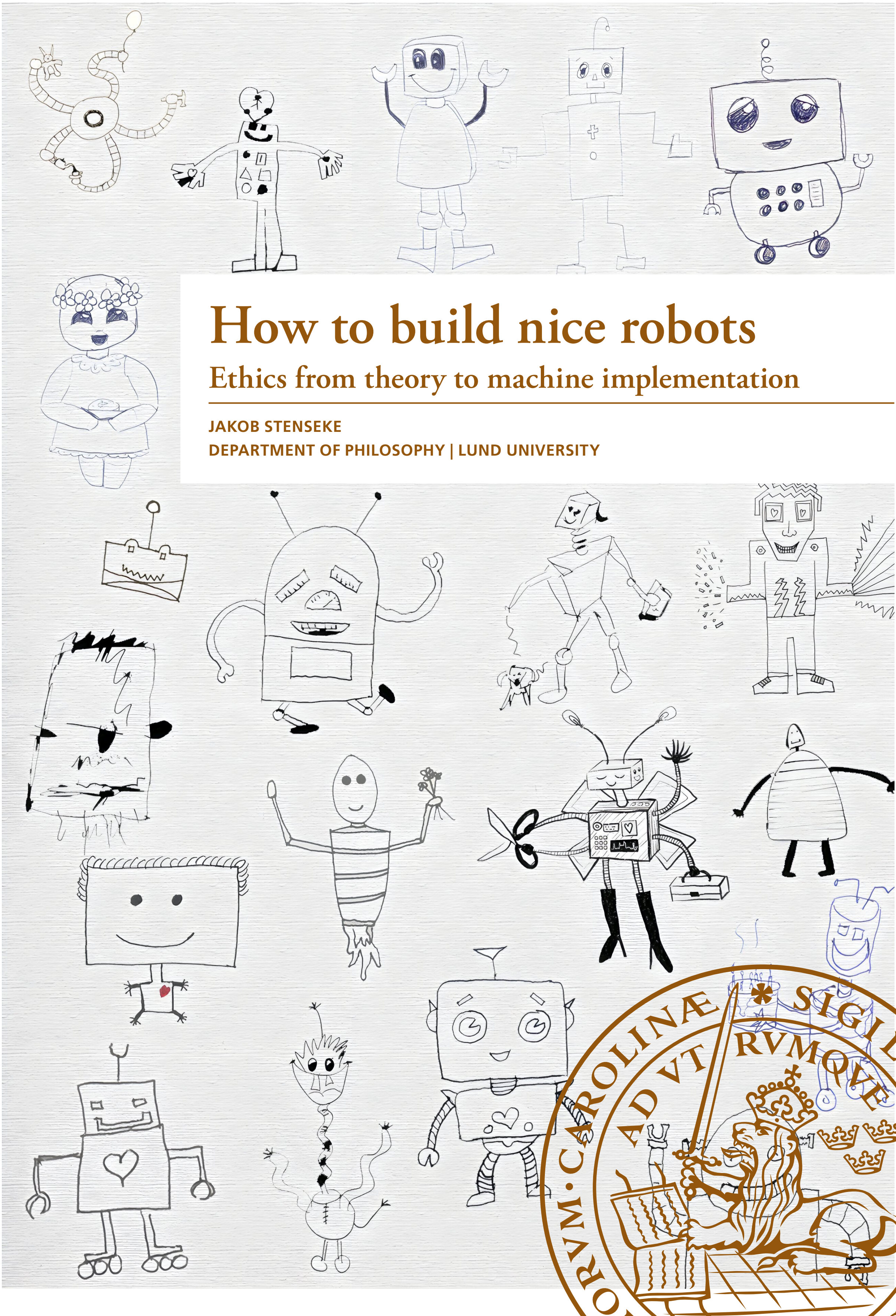
Ethics from theory to machine implementation

Stenseke, Jakob

2025

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00



How to build nice robots

Ethics from theory to machine implementation

JAKOB STENSEKE
DEPARTMENT OF PHILOSOPHY | LUND UNIVERSITY

How to build nice robots

How to build nice robots

Ethics from theory to machine implementation

Jakob Stenseke

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD)
at the Faculty of Humanities and Theology at Lund University to be publicly
defended on 26th of April at 10.00 in LUX C121, Helgonavägen 3, Lund

Faculty opponent
Vincent C. Müller

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation **Date of issue:** 2025-04-26

Author: Jakob Stenseke

Sponsoring organization: Wallenberg AI, Autonomous Systems and Software Program – Humanity and Socity

Title and subtitle: How to build nice robots: Ethics from theory to machine implementation

Abstract:

This thesis investigates morality from a computational perspective by examining how machines can be developed with capacities for moral reasoning and action.

It addresses how to overcome interdisciplinary boundaries between moral philosophy and computer science (Paper I), proposes a virtue-theoretic framework for artificial moral cognition (Papers II and III), and highlights issues of using normative ethics in moral machine design (Paper IV). Additionally, it analyzes how ethical decision-making is enabled and constrained by computational resources (Paper V) and explores artificial moral agency – first through an examination of Ishiguro's *Klara and the Sun* (Paper VI), and then by proposing a theory that bridges capacity-based and practice-based approaches (Paper VII).

The work unfolds along two main threads: Practically, it argues that moral machines should be developed 'bottom-up', with careful attention to the moral and non-moral aspects of the human practices in which they are meant to operate. Theoretically, it demonstrates that a computational approach to morality offers exciting opportunities to integrate diverse interdisciplinary insights, thereby enriching our understanding of morality itself.

Taken together, this work provides a smorgasbord of challenges and possibilities for moral machines, underscoring the need for interdisciplinary collaboration, technical feasibility, and grounding in human practice.

Key words: machine ethics, AI ethics, moral agency, interdisciplinarity, moral machines, artificial intelligence, virtue ethics, normative ethics, computational complexity

Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **Number of pages**: 322

ISSN and key title:

ISBN (print): 978-91-89874-90-9

ISBN (digital): 978-91-89874-91-6

Recipient's notes Price Security classification

Signature Date 2024-03-10

How to build nice robots

Ethics from theory to machine implementation

Jakob Stenseke

LUND
UNIVERSITY

Cover illustration: Jakob Stenseke (with the help of nice friends)

Faculty of Humanities and Theology

Department of Philosophy

Lund University

ISBN (print): 978-91-89874-90-9

ISBN (digital): 978-91-89874-91-6

Printed in Sweden by Media-Tryck, Lund University

Lund 2025

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Till Julia & Ormen

Table of Contents

Acknowledgements

How could there be nice robots without nice humans? This project has evolved in many ways over the years, but one thing has remained constant: the spectacular support I have received from so many people – from the research communities at Lund University, the WASP-HS network, academics spread around the world, and from friends and family. Tears of thankfulness form in my eyes when I realize how incredibly lucky I have been. I am deeply grateful to all of you.

I've had excellent supervision. Björn Petersson, thank you for your wise and steady guidance in matters large and small, academic and personal. Christian Balkenius, thank you for pulling me into this project and making me feel like anything is possible. Ylva von Gerber, thanks for many joyful conversations and for challenging me to think more carefully. And Trond Arild Tjøstheim, thanks for always leaving your door open, which has led to countless chats, far too many cups of coffee, and great friendship.

I've gotten lots of help on the way. Many have – via conversations, seminars, and feedback on drafts – been particularly helpful in bringing the content of this document into existence. For this, I owe special thanks to Asger Kirkeby-Hinrup, Niklas Dahl, Frits Gåvertsson, Andreas Stephens, Max Minden Ribeiro, Jiwon Kim, Amandus Krantz, Birger Johansson, Ingar Brinck, Alexander Velichkov, Shervin MirzaeiGhazi, and Alexander Tagesson.

I've had an amazing department. I'm probably the one who has benefited the most from its eclectic nature; recently coming from ventures in cognitive science, working in practical philosophy, yet sharing an office with theoretical philosophers (to Max, Andreas, Melina, Niklas, Fredrik, and Hubert – thanks for making the workplace feel like home). This has allowed me to constantly learn new things, bounce my thoughts off brilliant minds, and having a fantastic time doing so. I'm as thankful for your generosity as I am sorry for taking advantage of it. Thanks to Henrik Andersson, Olle Blomberg, Dan Egonsson, Anton Emilsson, Andrés Garcia, Mattias Gunnemyr, Anders Herlitz, Marta Johansson Werkmäster, Jiwon Kim, Marianna Leventi, Elsa Magnell, Jenny Magnusson, August Olsen, Robert Pál-Wallin, Erik Persson, Björn Petersson, Wlodek Rabinowicz, Paul Russell, Toni Rønnow-Rasmussen, Signe Savén, Martin Sjöberg, Daniel Telech, Patrick Todd, Jakob Werkmäster, Balder Ask Zaar, Mark Bowker, Hubert Hågemark, Carl-Johan Palmqvist, Melina Tsapos, Fredrik Österblom, Andrey Anikin, Thibault Boehly, Valentina Fantasia, Simon Grendeus, Agneta Gulz, Peter Gärdenfors, Magnus Haake, Jana Holsnova, Ivo Jacobs, Thomas Rejsenhus Jensen, Pierre Klintefors, Lona Lalic, Maybí Morell Ruiz, Jens Nirme, Helena Osvath, Mathias Osvath, Stephan Reber, Samantha Stedtler, Eva-Maria Ternblad, Matthew Tompkins, Betty Tärning, Gabriel Vogel, Annika Wallin, Anton Wrisberg, Daniel Zander, Agnès

Baehni, Vidar Bratt, Ellen Davidsson, Sarah Köglsperger, and Sara Midelf for making my journey such a delight.

I've had great support from heads and administrators. David Alm, Tobias Hansson Wahlberg, Petter Johansson, Martin Jönsson, Tomas Persson, Anna Cagnan Enhörning, Eva Sjöstrand, Annah Smedberg-Eivers, and Anna Östberg, thank you for all the administrative help and for keeping things running so smoothly.

A major highlight of this voyage has been our lively PhD community. Looking back, our time together feels like a perfect cocktail of fiery debates, unhinged silliness, and soul-nourishing camaraderie. I'm fortunate to count many of you not just as colleagues, but as close friends.

I've also been lucky in my ventures outside Lund. I am deeply grateful to the Wallenberg Foundations for their generous support, and to everyone involved in the WASP-HS network for organizing and participating in its courses, summer schools, and conferences. These gatherings have been a stimulating venue for discussing my work with like- and unlike-minded, and I cherish the new friendships and perspectives it has brought into my life.

Special thanks to Jakob Hohwy and everyone at M3CS at Monash University, as well as Pat, Kelly, Fen, Erin, Sue, and T-Bone for making my 2024 spring semester in Australia an absolute delight. Thanks to Catrin Misselhorn, Thilo Hagendorff, Felix Lindner, and Thomas Metzinger for stimulating conversations during my 'German Machine Ethics Trip' in 2022 (and to Erasmus+ for financing it!). Thanks also to everyone else I have met at along the way for influencing my work in constructive ways, including (but not limited to) Sven Nyholm (my formidable mock opponent), Per Austrin, Olle Häggström, Björn Lundgren, Nick Barrow, Anders Søgaard, Ajay Vishwanath, Dorna Behdadi, Paul de Font-Reaulx, and Marc Champagne.

Some have been massively important to me in past and present chapters of my life. The Mayfield Street Elite, Zekes boys, Amsterdam peeps, Feelium gang, the Stenseke, Johansson, Wright, and Nelson families, Simon Folger, Gustav Stenseke, Francis Patrick Brady, Gustav Törnblom, Malte Grahm, Jeroen Mutsaers, Eoin Ó Cuinneagáin, Thomasine Eriksson, Martin Skarbäck, Philip Højme, Jonna Restin, Karl Samson, Adel Al Khozaae, Mahyar Hemmati, Mariam Marlen Mirström – you all have a special place in my heart and mind.

To my family, Irén, Lars, Oskar, Alfred. I am thankful for the unwavering love and support you have given me throughout all my life.

Finally, there is one person who stands out above all. Julia Wright, thank you for treating me like a like a dog – feeding me, taking me on walks, listening to my howls, and showing me unconditional love and care, through sunshine and storms. You are my greatest joy. ILYSFM.

Abstract

This thesis investigates morality from a computational perspective by examining how machines can be developed with capacities for moral reasoning and action.

It addresses how to overcome interdisciplinary boundaries between moral philosophy and computer science (Paper I), proposes a virtue-theoretic framework for artificial moral cognition (Papers II and III), and highlights issues of using normative ethics in moral machine design (Paper IV). Additionally, it analyzes how ethical decision-making is enabled and constrained by computational resources (Paper V) and explores artificial moral agency – first through an examination of Ishiguro's Klara and the Sun (Paper VI), and then by proposing a theory that bridges capacity-based and practice-based approaches (Paper VII).

The work unfolds along two main threads: Practically, it argues that moral machines should be developed 'bottom-up', with careful attention to the moral and non-moral aspects of the human practices in which they are meant to operate. Theoretically, it demonstrates that a computational approach to morality offers exciting opportunities to integrate diverse interdisciplinary insights, thereby enriching our understanding of morality itself.

Taken together, this work provides a smorgasbord of challenges and possibilities for moral machines, underscoring the need for interdisciplinary collaboration, technical feasibility, and grounding in human practice.

List of papers

This thesis is based on the following papers, referred to by their Roman numerals:

Paper I

Stenseke, J. (2022). Interdisciplinary confusion and resolution in the context of moral machines. *Science and Engineering Ethics*, *28*(3), 24.

Paper II

Stenseke, J. (2023). Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*, *38*(4).

Paper III

Stenseke, J. (2024). Artificial virtuous agents in a multi-agent tragedy of the commons. *AI & SOCIETY*, *39*(3).

Paper IV

Stenseke, J. (2023). The use and abuse of normative ethics for moral machines. In *Social Robots in Social Institutions*, R. Hakli, P. Mäkelä, J. Seibt (Eds.), pp. 155-164, IOS Press.

Paper V

Stenseke, J. (2024). On the computational complexity of ethics: moral tractability for minds and machines. *Artificial Intelligence Review*, *57*(4), 105.

Paper VI

Stenseke, J. (2022). The morality of artificial friends in Ishiguro's *Klara and the Sun*. *Journal of Science Fiction and Philosophy Vol.5*.

Paper VII

Stenseke, J. (Unpublished manuscript). Knowing and owing each-other: on the co-construction of moral agency across time and space.

List of publications not included in the thesis

Kirkeby-Hinrup, A. & Stenseke, J. (Forthcoming). The psychology of LLM interactions: the uncanny valley, and other minds. *Journal of Psychology and AI.*

Stenseke, J. (2025). Optipolitis: is politics an optimization problem? In *Social Robots with AI: Prospects, Risks, and Responsible Methods*, J. Seibt, P. Fazekas, O. Santiago Quick (Eds.), IOS Press.

Stenseke, J. & Tagesson, A. (2025). The prospect of artificial empathy: a question of attitude? In *Social Robots with AI: Prospects, Risks, and Responsible Methods*, J. Seibt, P. Fazekas, O. Santiago Quick (Eds.), IOS Press.

Kirkeby-Hinrup, A., Stenseke, J. & Overgaard, M. (2025). Evaluating the explanatory power of the Conscious Turing Machine. *Consciousness and Cognition, 124,* 103736.

Tagesson, A. & Stenseke, J. (2024). Do you feel like (A)I feel? *Frontiers in Psychology.*

MirzaeiGhazi, S. & Stenseke, J. (2024). Responsibility before freedom: closing the responsibility gaps for autonomous machines. *AI and Ethics.*

Stenseke, J. & Balkenius, C. (2022). Assessing the Time Efficiency of Ethical Algorithms. *Proceedings of the 8th International Workshop on Intelligence and Cognition (AIC 2022)*, *CEUR-WS.*

Stenseke, J. (2021). Persistent homology and the shape of evolutionary games. *Journal of Theoretical Biology*, *531*, 110903.

Chapter 1 – Introduction

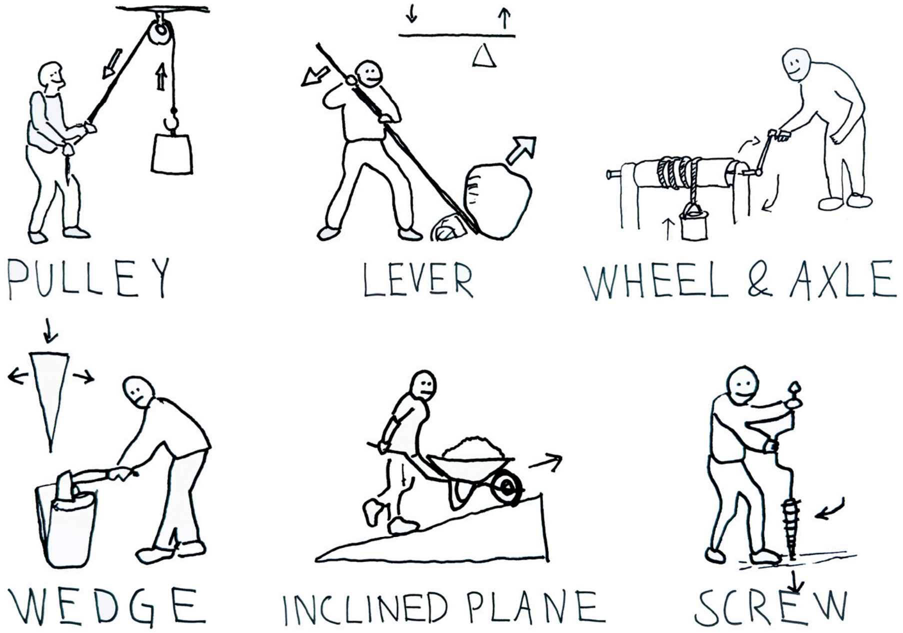


Figure 1: Simple machines

1.1. Robots

The title of this thesis is a bit misleading. Surely, it is about nice robots and how to build them. But with ‘robots’, I have a rather large collection of various machines in mind. Chances are that some members of this collection may not be the same as the kind of robots you think of when you think of robots.

The robots I have in mind are all computational systems, machines that can be programmed to carry out operations automatically. They may perform tasks related to physical labor, like pushing and carrying boxes around a warehouse. Other robots may be specialized for mental work, such as solving math problems or memorizing the right pathway through a maze. Some may be mechanically simple, but most are quite advanced, potentially utilizing billions of transistors in modern microchips and running cutting-edge software. Typically, robots are to some extent capable of doing some things autonomously, without human supervision or control. Nowadays, many robots can learn things – from examples or by trial-and-error – while others are restricted to following static code.

Another classic feature of robots is that they have a physical body, often with features resembling animal or human forms, with legs for walking, arms for grabbing, and a face for talking. But the robots I will talk about might as well be best known for their existence in the digital realm; as software programs, chat bots, or characters in some virtual world, with or without a distinct avatar to signify their presence.

I will not be picky with the definition of 'robot'. There are countless variations of robots, and it is surprisingly difficult to identify sharp boundaries of what differentiate them from non-robot machines and "AI systems", the latter referring to machines exhibiting some form of intelligence.

One reason for this terminological difficulty is that most robots are combinations of the non-robotic machines that preceded them, just as their programs may employ any of the oldest or latest methods in artificial intelligence.

Once upon a time, however, there were only six machines: the pulley, the lever, the wheel and axle, the wedge, the inclined plane, and the screw (Figure 1). For Renaissance scientists, these so-called *simple machines* were thought of as elementary building blocks out of which all other machines could be constructed.[1] Echoes of this mechanistic simplicity can still be felt today. If you disassemble a bicycle, you will find variations of levers, pulleys, and wheels. Not too long after the Renaissance, the great variety of sophisticated machines developed during the Industrial Revolution made it impossible to describe and analyze machines using the six basic categories. Today, as we interact with the convenient graphical user-interfaces of computer software and smartphone apps, we never even see the mechanisms under the hood. And if you take apart a modern robot, you will likely find more than levers, pulleys, and wheels.

Another reason for the terminological difficulty is that our concepts of robots, like artificial intelligence, are moving targets. The mechanical automata of past centuries may to us seem more like puppets if put next to the robots of the 21st century (Law, 1997; Truitt, 2015). The "good old fashioned" AI methods discussed at the 1956 Dartmouth Summer Research Project on Artificial Intelligence – often thought of as

[1] The concept of simple machines is often said to originate with Archimedes around the 3rd century BC, who discovered the mechanical advantage of the lever. Later on, Galileo Galilei (ca. 1600, in *Le Meccaniche*) identified the underlying mathematics of simple machines in terms of force amplifiers (Cardwell, 2001; Usher, 1954).

the founding event for AI as a distinct field – may have little in common with the artificial neural networks and machine learning algorithms closely associated with – and often synonymous to – AI today (Nilsson, 2009; Russell & Norvig, 2020). In this way, AI and robots can sometimes refer to whatever is hot or next up on the technological frontier; and innovations that once were considered "cutting-edge AI" may, as they become widely incorporated into general applications, no longer be called AI.

But vague concepts have certain advantages over precise ones. They can dynamically connote a variety of ideas stretching across time and space, constrained only by the limits of imagination and association. In this sense, 'robot' is flexible enough to capture automata of the past, present, and future.

What I like the most about robots are their relationship to the cultural zeitgeist. What they can symbolize for us. It is not a coincidence that the most famous robots – e.g. C-3PO, Terminator, or HAL 9000 – all happen to be fictional. Robots can symbolize a new kind of being, perhaps created in our own image; or one that is completely alien. It can be a perfect being, devoid of human flaws. It can also be something from apocalyptic nightmares – something that will come to take over the world or destroy it altogether. The kind of robot I will imagine is a nice one.

1.2. Nice? In what way and for whom?

There is another sense in which the title of this thesis can be misleading. Just as there are many kinds of robots, machines, and AIs, there are of course many ways in which something can be nice.

Most machines are nice, at least in the sense that they do what they are supposed to do. A nice car can drive us to remote places. A nice washing machine cleans our clothes. Typically, they are nice *for* something, like achieving some goal, and *for* someone – say, being nice for humans who want to go to remote places or have clean clothes.

But all machines are not nice, and no machine is all nice. Some of them, like guns or autonomous weapon systems, can be used to do unpleasant things. Factory machines allow human societies to produce more than what they need, creating enormous stress on the natural environment.

There are now several subfields of research exploring various ways to ensure that machines – including AI systems and robots – remain nice for individuals, societies, and the environment.[2] Nice machines should be *safe* – prevented from being misused or causing harmful outcomes (AI Safety). The behavior and inner workings of nice machines – particularly those with sophisticated capacities for learning and reasoning – should be *transparent* and easy to *explain* (Transparent and Explainable AI). Nice machines that aid decision-making should be *fair* so as to not discriminate against certain groups of people (Algorithmic Fairness). And nice machines should be developed and used in *responsible* ways (Responsible AI), and sensitive to social, economic, and environmental *sustainability* (Sustainable AI).[3]

But the kind of nice machines that I will explore are those that have some form of niceness built into them. They are not merely nice *for* something or *someone*, but also nice themselves. Surely, it would be great if these machines were also safe, transparent, explainable, fair, responsible, and sustainable in the ways just described. But this is not the main focus. As the subtitle indicates, by niceness, I am really referring to morality and ethics, as in machines that are able to reason or act based on some conception of what is ethically right and wrong, morally good and bad.

Replacing "nice" with "moral", however, does not clarify the issue much. After all, people have wrestled with the concept of morality since the dawn of time. And although various philosophers and prophets have offered interesting answers over the ages, disagreements on the nature of morality remain as prevalent as ever. But don't worry; a significant chunk of this project consists of trying to make it clearer what morality means – and what it could mean – in the context of machines.

1.3. Ten meters from the robot lab

Back when I began this project in 2020, I was more interested in building and less interested in thinking about what morality means, being fed up with certain philosophical debates – about morality, consciousness, etc. – that seemed to go in circles. I was eager to get my hands dirty in the robot lab, conveniently located just ten meters from my office. My attitude was that of an engineer: the best way to learn about something is to try to build it. Four years later, and I haven't (yet) made it to the robot lab. To build something, you first need some kind of blueprint. As Kurt

[2] See Huang et al. (2022) for a brief exposition of AI ethics; Coeckelbergh (2020) and Boddington (2023) for two longer overviews.

[3] See, e.g., Amodei et al. (2016) for issues in AI safety, Barredo Arrieta et al. (2020); Ehsan et al. (2021); Larsson and Heintz (2020) for transparency and explainability of AI, Mitchell et al. (2021) for algorithmic fairness, Dignum (2019) for responsible AI, and Van Wynsberghe (2021) for sustainable AI.

Lewin once said, “there is nothing more practical than a good theory” (Axelrad, 1951). So, over time my work became less about building nice robots and more about finding good theories of niceness that can act as blueprints for building nice robots. Ironically, this has led me back to some of the philosophical debates – about morality, consciousness, etc. – I once thought I had managed to escape.[4]

By the time of writing this, I am probably further away from the robot lab (metaphorically speaking) than where I was four years ago. Yet, in this journey, I have also discovered many fascinating theories that have helped me in my quest for a blueprint. And in this work, I will tell you about some of them.

This leads us to the last potentially misleading term in the title: “build”. I will not give anything that resembles easy-to-follow IKEA-instructions of robotic parts that could be assembled in this or that way. Nor will I, like the Renaissance scientists, present a collection of elementary nice machines of which all nice robots can be built.

That being said, some of the work provides recipes – ranging from detailed descriptions of algorithms and AI methods to more abstract computational architectures and frameworks – that can support the construction of nice robots of various sorts (Papers II, III, & V).

Other parts of the build plan consist of more philosophical inquiries, e.g., on how aspects of morality could or should be understood from a computational perspective (Papers I & V), and what a moral agent is and whether a robot could be one (Papers VI & VII). Yet other parts of the build plan merely present important things to consider – such as technical and normative constraints – before one even begins to build, or during the process of building, a nice robot (Papers I, IV, & V).

It should be emphasized that I will not articulate a specific overarching argument or coherent vision. Rather, the work should be seen more as a recipe book, containing a selection of things – considerations, issues, frameworks, and results – that in various ways are important for those who want to increase their chances of one day making it to the robot lab.

[4] As a postscript remark, this circuitous journey is reflected in the papers themselves. As a commentator noted on an earlier draft of the thesis, there is something of a plot twist occurring between Papers I-III and IV-VII: where the former seems more optimistic about the prospects of getting to the lab, the latter adopt a more skeptical stance and get further entangled in convoluted issues. Consequently, some points raised in the later papers – particularly IV and VII – could in fact be leveled as critiques of arguments presented in the earlier papers. I hope that readers will approach this work with a generosity of interpretation that takes the overall contribution and trajectory into account, rather than fixating on its internal inconsistencies.

1.4. Time flies in the age of machines

One reason for this limited yet eclectic scope is that the horizon of moral machines is constantly expanding and changing direction in unpredictable ways. In 2020, there was already a surge of literature on the ethics of AI under umbrellas such as AI Ethics and AI Safety. And over the years, additional umbrellas have not only successfully established themselves in the academic zeitgeist, but some efforts have worked their way into legalization. A good example of this is the European Union Artificial Intelligence Act (AI Act), a legal and regulatory framework that came into force on 1st of August, 2024 (European Commission, 2024). Yet, it should be noted that this ongoing boom is only a reaction to the widespread deployment of modern AI technologies in the past few years, which in turn stems from the extraordinary advancements AI has undergone in the past two decades. In 2010, the total number of AI publications was roughly 88,000; in 2022, it was over 240,000 (Perrault & Clark, 2024).

I assume that teachers around the world may already be planning on how to secure the quality of education for the next term, given that easy-to-use Large Language Models such as ChatGPT, with hard-to-detect capacities for plagiarism, are one step away from students' fingertips (Farazouli et al., 2024). Likewise, due to the spread of misinformation, hallucinations, and deepfakes propagated by generative AI, some of us may struggle to tell real from fabricated content (Monteith et al., 2024). In short, AI technologies are bound to – if they haven't already – affect or even transform most domains of human life. As a result, the public awareness of the ethical problems of AI have skyrocketed beyond what I could have imagined.

A fortunate upshot is that most people I interact with recognize the relevance of my dissertation work – and are often eager to talk about it – without my needing to motivate its importance. As a researcher, this is a true luxury. A less fortunate consequence is that it becomes impossible to stay properly informed on all fronts of AI development, deployment, and impact. An even more unfortunate consequence is the risk that some of the things I address will become outdated even before this thesis hits the press, or simply be lost in the vast ocean of AI buzz. While this has to some extent informed my choice of topics to look into, I can only hope that some of it will stand the test of time, and if not, at least for some period of time. Unless humanity embarks on a Butlerian Jihad,[5] there will eventually be much better ways to build nice robots than what will be recommended here. I can only hope that this

[5] In the *Dune* series by Frank Herbert, the Butlerian Jihad refers to a conflict that resulted in the total destruction of "computers, thinking machines, and conscious robots" (Herbert, Dune, 1965, Terminology of the Imperium: Jihad, Butlerian). The event is named after Samuel Butler, who warned about the apocalyptic dangers of thinking machines already in 1863 (Butler, 1863).

work can play some part in getting there, but I recognize that it might as well take a completely different path.

1.5. The road to the lab

Papers I-VII constitute the bulk of the thesis. Therefore, the primary aim of Chapters 2–6 is to provide context for the papers and demonstrate how they cover some aspects of building moral robots. As a secondary aim, I will also use these chapters as an opportunity to expand on some ideas not covered at length in any of the papers.

Chapter 2 introduces the field of *machine ethics* along with its central research questions, covering the *why*, *can*, and *how* of nice robots.

Chapter 3 describes three grand challenges for building moral machines, namely, that morality is multifaceted, contentious, and hard. The chapter also serves to justify the mix of methods and disciplines employed in the project.

Chapter 4 explores how standards of right and wrong can provide recipes for the construction of nice robots. The chapter identifies some problems with this methodological strategy and provides ideas on how to overcome them through the notion of *convergence*.

Chapter 5 turns to more philosophical issues about moral agency, asking whether machines can *really be* nice. It describes three capacities that are central to standard conceptions of moral agency – namely rationality, autonomy, and consciousness – and discusses whether and in what way AI systems of today and tomorrow can have them. The chapter then situates the moral agency capacities alongside alternative approaches to moral agency and presents a theory that seeks to reconcile them.

Finally, Chapter 6 summarizes the thesis, offers a conclusion that synthesizes the project into two main threads, and gives an exposition of the papers.

Chapter 2 – Machine ethics

Machine ethics is an interdisciplinary field at the intersection of artificial intelligence (AI), philosophy, and cognitive science (along with related disciplines). Its central focus is on creating Artificial Moral Agents (AMAs), i.e., machines imbued with moral capacities, such as moral reasoning and ethical decision-making.[6]

The field can be seen as a close associate to other strands of AI ethics mentioned in the introduction, such as AI Safety, Responsible AI, Explainable AI, and Algorithmic Fairness. What these strands have in common is that they, in some way or another, tackle the ethical issues of AI. Yet, what makes machine ethics different from other efforts is that it envisions the development of explicitly *moral* machines *as* a possible – or even reasonable – pathway to address some of the ethical issues of AI.

The field can be further organized along three central research questions, exploring the normative desirability (why), the theoretical feasibility (can), and the technical engineering (how) of AMAs. Here, I will elaborate on how the *why*, *can*, *how* have been addressed within machine ethics, which also serves as a background for presenting my own contributions.

2.1. Why do we want nice robots?

Before one even begins to build a nice robot, it is important to have some reasonable answer to *why* one wants to do so. After all, human societies seem to have been doing fine without them, so why are they needed now? What are the benefits of creating them, and how do these benefits weigh against the potential drawbacks?

The prevailing view in the wider public seems to be negative. Machines may help to drive us and keep our clothes clean, but there is a strong reluctance to entrust

[6] As stated by machine ethicists Michael and Susan Anderson, "the ultimate goal of machine ethics is to create autonomous ethical machines" (Anderson & Anderson, 2007, p. 15). For three accessible introductions and overviews of machine ethics, see Anderson and Anderson (2011); Pereira and Lopes (2020); Wallach and Allen (2008).

them with moral decision-making. That is better left for people themselves. For instance, a large study by Bigman and Gray (2018) found that participants were averse to machines making morally-relevant decisions in medical, legal, and military contexts, and that this aversion persisted even in cases where the machine-made decisions had positive outcomes.

This observation should push the enthusiastic nice-robot-builder to better motivate their development. For even if they were successful in creating moral machines, the aversion towards machine morality would stop any potential consumers from buying and using them. To this end, a more comprehensive exposition have been provided by Van Wynsberghe and Robbins (2019), who critically analyze the reasons commonly invoked as justification for constructing moral machines. Here I will briefly discuss five of them:[7]

(i) *Moral machines are inevitable* – The first justification is that the creation of morally capable machines is in some sense inevitable.[8] It starts by extrapolating from observations of the ubiquitous deployment of AI and autonomous machines in morally salient contexts; for instance, the growing number of self-driving cars on public roads, or the variety of intelligent systems being employed in medicine, education, and elderly care. A common example is that of accident-management for autonomous vehicles, which have been argued to present moral dilemmas (Keeling, 2020; Nyholm & Smids, 2016). There is also reason to suspect that these morally salient contexts are prevalent. As Scheutz writes, "any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation, where the artificial agent finds itself presented with a moral dilemma where any choice of action (or inaction) can potentially cause harm to other agents" (Scheutz, 2016, p. 516). Now, if increasingly capable machines take on increasingly complex roles in human society – as chauffeurs, teachers, and doctors – it is reasonable to expect them to adhere to the moral standards associated with those

[7] There are two additional reasons discussed by Van Wynsberghe and Robbins that are omitted here, which revolve around *complexity* and *increasing public trust* for AI and autonomous machines. The first is the idea that the behavior of sufficiently complex AI systems will become so unpredictable that they need to "have 'ethical subroutines' of their own" (Allen et al., 2006, p. 14). The second reason is that, if machines can be morally competent, it will increase our trust and confidence in these systems to act autonomously on our behalf. Against this, Van Wynsberghe and Robbins pinpoint to an inconsistency between promoting the development of moral machines for reasons of *trust* and reasons of *complexity*: moral machine developers cannot simultaneously argue that machines need to be moral due to the increased complexity and unpredictability of their functioning *yet* expect us to increase our trust for these systems (as unpredictability conflicts with trust).

[8] See, e.g., Allen and Wallach (2012); Anderson and Anderson (2010); Scheutz (2016); Wallach (2008) for different variations of this claim.

roles. Ergo, it is necessary to imbue machines with capacities for making ethically informed decisions.

(ii) *Prevention of human harm* – The second justification is straightforward: the development of AMAs will serve to prevent or reduce the potential harms that machines can inflict on humans. The rationale is that machines are demonstrably capable of causing harm to humans, and these harms can be mitigated or reduced by making the machines morally competent. In short, a moral machine will cause less harm than what a non-moral machine would do. A classic example of this is the first of the Three Laws of Robotics, as described by science fiction author Isaac Asimov: "A robot may not injure a human being or, through inaction, allow a human being to come to harm" (Asimov, 1950, p. 40).[9]

(iii) *Prevention of immoral usage* – The third justification is also fairly straightforward: moral machines, compared to amoral machines, are less likely to be misused, e.g., for malicious human purposes. For example, while an amoral robot might assist a burglar in breaking into a house, a moral robot would not; it would prevent itself from being used to facilitate such an enterprise.

(iv) *Moral superiority* – The fourth justification is that moral machines have the potential to be morally superior to humans, e.g., being more capable as moral reasoners and devoid of human flaws. As computer scientist James Gips wrote 30 years ago, "not many human beings live their lives flawlessly as moral saints. But a robot could." (Gips, 1994, p. 250).[10] For instance, one might point to various examples of human frailty – her emotional biases, short-sightedness, and limited cognitive capacities – and compare these with instances where AI has achieved super-human performance (e.g., Chess, Go, protein folding), and then ask: which is more capable of making consistent, rational, and impartial decisions? An example of this justification in the military context has been given by Arkin (2007), who argues that autonomous military robots would, if they were programmed to follow the Laws of Just war, not pillage, murder, or rape the civilians in the villages conquered during warfare.

(v) *To better understand morality* – The fifth justification is that building machines with moral capacities will lead to a greater understanding of human morality. This is captured in the quote "What I cannot create, I do not understand", attributed to

[9] Indeed, by illustrating the challenges and complexities of implementing ethics into machines through a series of short stories, Asimov's may have been the first machine ethicists.

[10] See also Dietrich (2001) for a more rigorous defense of this idea.

physicist Richard Feynman.[11] The idea is that, when faced with the task of creating a moral machine, we learn about various background assumptions implicit in the ethical theories themselves (Gips, 1994). Even seemingly simple ethical principles may presuppose quite sophisticated cognitive machineries. For instance, to adhere to the "do no harm"-principle, an agent needs the appropriate cognitive capacities and knowledge to understand what "harm" means, and the types of actions and states of affairs that would (and would not) constitute harm. As such, moral machines offer a valuable opportunity to "reverse engineer" morality; to computationally model or recreate what we consider to be the most relevant or essential aspects of human morality. A relevant analogy is that of teaching something to another person. Understanding something is one thing, but understanding it well enough to teach it requires a deeper form of understanding. Teaching it to a computer is yet another matter, as it pushes one to formalize the understanding as an algorithm. It is in this way that developing a moral machine compels us to articulate morality sufficiently well and clearly to implement it computationally. While it is unlikely that a computational model could capture all relevant aspects at the first attempt, it is nonetheless an artifact that can be discussed and improved upon. In this process, it is possible that we get insights into what morality is and how it works; insights that would have been missed if we only adhered to the traditional (human) approaches.

The moral-machine-building enthusiast may invoke some or all of the points (i)–(v) described above and rest assured: their project is justified. However, Van Wynsberghe and Robbins (2019) ultimately conclude that none of the reasons stands up under closer scrutiny. Against (i), they argue that, although machines will inevitably play a central role in many morally salient contexts, it is simply *not* inevitable that they should be delegated any significant moral role in those contexts. After all, these machines need not be *moral* machines in order to perform their function well in morally charged contexts. A heart monitor does not need the ability to reflect upon right and wrong in order to decide its next course of action; it should simply report information to the doctors that will make any morally relevant calls.

Along similar lines, Van Wynsberghe and Robbins argue against (ii) that any prevention or reduction of harm that moral machines could attain may equally well be achieved by simply focusing on developing *safe* machines.

In response to (iii), they point out that preventing certain uses of machines (e.g., malicious ones) comes at the cost of reducing human autonomy and the ability of humans to override machine decisions they, for various reasons, deem erroneous. We may think of a breathalyzer that prevents someone from using their car to escape

[11] Richard Feynman, who received the Nobel prize in physics for his work on quantum electrodynamics, purportedly had the quote written on his blackboard at the time of his death (Way, 2017).

violence, or a robot servant that refuses to fetch another can of beer as that contributes to poor health.

Van Wynsberghe and Robbins goes on to contest (iv) – the moral superiority of machines – by analyzing the controversial idea it presupposes: that there are measures of what constitutes moral superiority. Certainly, a calculator's arithmetic capacities may exceed those of a human. But can one person be morally superior to another *tout court*? If so, this assumes the existence of standards of morality by which we can evaluate moral superiority, and furthermore, that we can have knowledge of these standards. Of course, far from everyone agrees with these assumptions (more on this and other disagreements in 2.4.2.). Here, proponents of machine morality might argue that machines – free from human bias and short-sightedness, and possessing vastly superior cognitive capacities – could surpass humans in discerning these moral standards. We can imagine a scenario in which a moral machine comes to us and says: "I have discovered a moral truth to which everyone should adhere, a truth that no human could discover due to their inherent limitations". But would we listen? Problematically, adhering to the machine's discovery seems to require that we take a leap of faith, blindly accepting their asserted superiority.

But the problem does not stop there. Even *if* we were to blindly accept the moral superiority of machines, one might ask: what consequences would this have for human morality? Consider the recent proliferation of generative AI tools in areas such as programming and graphical design. Instead of actively practicing coding and design skills, many programmers and designers now prompt generative AI to perform these tasks for them. In these contexts, while the end-result may be satisfactory – the generated code functioning as intended, and the generated design conveying the desired artistic message – the process leads to human deskilling in programming and graphic design. Now, what would such outsourcing entail in moral contexts? Along these lines, Vallor (2016) has argued that outsourcing moral decisions to machines may lead to undesirable forms of moral deskilling in human beings. The concern is that if we do not actively cultivate our capacity for moral reasoning and imagination – traits traditionally considered critical for leading a fulfilling life – there is a risk that we shrink the opportunities for human flourishing and lose the ability to envision a future worth wanting.

Finally, with respect to (v) – the idea that building moral machines will lead us to a greater understanding of human morality – Van Wynsberghe and Robbins gives the following consideration:

> […] [E]thical theories are not (and have little to do with) how people reason morally so the work doesn't help understand *human* morality. […] [H]uman morality, in the descriptive sense, is dependent upon many complex factors and building a machine that tries to perfectly

> emulate morality must use each of these factors combined rather than rely on ethical theory alone. (p. 731)

Weighing up the arguments for and against, is the moral machine-building project justified? I find merit in all of the five discussed justifications. Instilling some capacity for ethical reasoning in the design of machines may not be inevitable, but it may very well be desirable, particularly if it serves to prevent or reduce harm and immoral usage. At the same time, I concur with Van Wynsberghe and Robbins' critique. Some of the benefits comes with problematic drawbacks and costs, such as shrinking the space of human autonomy and opportunities for practicing moral skill (iii & iv). Moreover, several benefits may also be achievable without attempting to instill some form of 'ethical subroutine' in a machine (ii, iii, iv, & v). For instance, rather than implementing an ethical module in a self-driving car, we could focus on making it *safe,* so as to reduce the likelihood of the car ever encountering morally salient situations (and be forced to make potentially objectionable decisions). Rather than deferring to morally superior machines, do we not still have much to learn and discuss before we could settle what such 'superiority' would entail (iv)? And how much can we learn about morality from developing moral machines in comparison to what we can learn about human morality from studying humans (v)?

Instead of abandoning the project already at this juncture, I will offer two points that I believe can serve to justify its continuation.

The first point concerns the definition of a "moral machine". As indicated by their response to (v), the sort of moral machines Van Wynsberghe and Robbins appear to criticize are those that are *only* based on ethical theory, neglecting the "many complex factors" upon which human morality is (allegedly) dependent. As we will see later on, this is indeed a common methodological strategy for moral machine builders: developing algorithms that follow the prescriptions of some normative theory, while overlooking the complex factors – e.g., cognition and resources – that are necessary for the successful application and justification of that normative theory (Papers IV & V). And in the rest of this chapter, as well as in Chapters 3 and 4, I will develop additional criticisms of this methodology (and ideas on how to overcome them). For now, it remains an open question whether Van Wynsberghe and Robbins' critique leaves room for moral machine projects that seek to do justice to the "many complex factors" of human morality, and in particular those motivated by the quest to better understand it. Indeed, a primary aim of this thesis is to do justice to at least *some* of these complex factors and demonstrate how moral machines offers a new avenue to illuminate them.

The second point starts by highlighting a peculiar circularity: developing moral machines *is* a necessary step in properly assessing their desirability. In other words, we do not seem to have any clear insight into the *why* of nice robots before we have more concrete ideas about the kind of nice robots we *can* build.

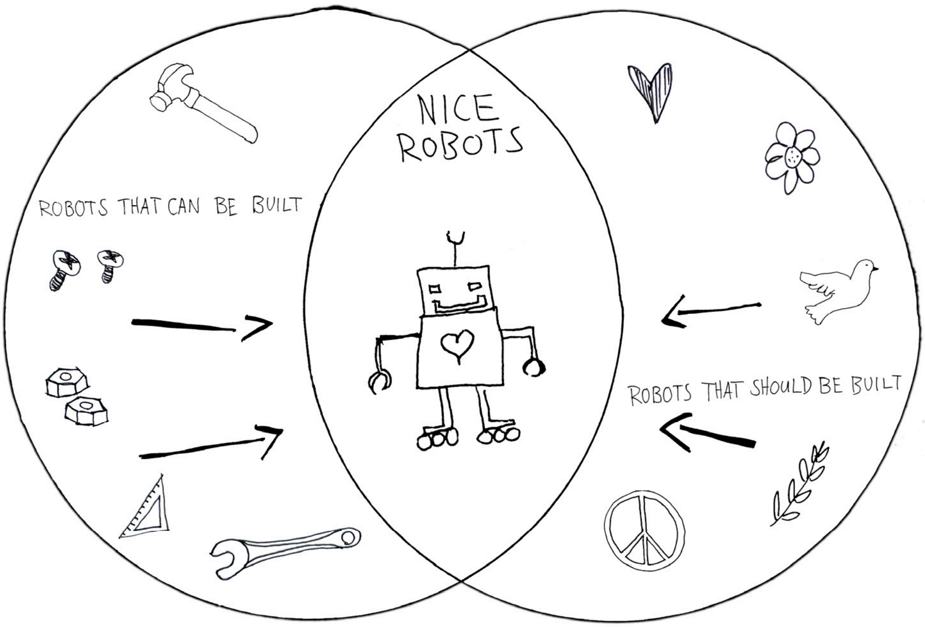


Figure 2: The mutual constraints of cans and oughts

This presents a conflict that has resided in me throughout this project: between building and critically thinking about moral machines. On the one hand, compelling reasons must exist for attempting to build such machines; yet, on the other hand, what those machines really are like – and could be like – may remain abstract and unclear without concrete examples.

This inner conflict resulted in Paper I, titled “Interdisciplinary confusion and resolution in the context of moral machines”. The paper explores the tension between those who simply want to *build* without asking too many whys – in particular, the engineers and computer scientists eager to build moral machines – and those who *criticize* the buildings (or the reasons for building them), sometimes with limited insight into the craft of building.

As you might have guessed, there are obvious problems with pursuing only one of these paths. But the tension between them runs surprisingly deep, as there are practices, concepts, and aims inherent to the disciplinary perspectives that lead one to take one over the other. Problematically, this tension facilitates incommensurable views on the prospect of machine morality, which makes it difficult for builders and thinkers to work together and find synergies that utilize their respective strengths.

One of the remedies proposed in Paper I is to use the technical *cans* and moral *oughts* of moral machines as mutual constraints (Figure 3). Those focusing on the *oughts* can criticize a range of possible moral machines, but their critiques are ineffective unless it is grounded in technical feasibility or addresses a specific technological artefact. Conversely, without asking *why*, builders might go on to deploy moral machines in real-world environments even in the absence of adequate justification for their need or desirability. Without the appropriate tools to talk about it, there is also a risk of that the builder and thinker fundamentally misunderstand each-other. The ethicist may criticize an artefact that only exists in their fantasy, while the builder's idea of a moral machine fails to satisfy essential criteria of what morality – according to non-builders – is.

The point is not meant to generalize beyond moral machines, implying that one must first invent some X before knowing whether it is a good idea to invent X. Some inventions may be so dangerous that even the idea of them is sufficient grounds to argue against their development. Nor does it imply that moral machine projects are, at any stage of development, exempt from justifying their *why*s. And as we will see in Chapter 5, certain machine projects come with significant problems and risks.

The point is rather that, in the context of moral machines, whys, cans, and hows are best pursued in unison. Yes, there are indeed valid reasons for creating moral machines. And yes, there are valid reasons to be cautious and skeptical. Presently, however, the question of whether the creation of nice robots is justified remains to be illuminated by considered whether, how, and to what we extent they can be moral. And this, I hope, serves as sufficient justification to look into *can* and *how.*

2.2. Can robots be nice, really?

The *can* concerns the theoretical possibility of creating moral machines. A skeptic might say, surely robots cannot feel, think, or will in the appropriate ways that would qualify them as moral beings. To this, the eager nice-robot-builder responds: while it is true that no robot possesses these capacities today, what stops them from acquiring them in, say, 100 years? In addition, the builder shows various graphs demonstrating continuous increases in the performance of AI systems over a broad range of cognitive tasks.

Against the future possibility-consideration, the skeptic presents John Searle's well-known Chinese Room argument (Searle, 1980). Imagine sitting in a room where you follow instructions written in English to manipulate Chinese symbols. By following the rules, you can produce adequate responses in Chinese without *understanding* the language. To an outside observer, however, it appears as if you do understand Chinese. Analogously, a machine could produce outputs and behaviors that appears to demonstrate moral understanding. But it is merely following instructions and does not truly understand it, in the same way you don't truly understand Chinese by following instructions inside the room. And I, the skeptic says, believe the same will be true for machines even in 100 years.

The nice-robot-builder responds: your thought-experiment seems to rely on rule-following. That may capture the "good old fashioned" AI methods of the 60's, which centered around the use of logical inferences and hard-coded symbolic representations of the meaning of words. But the deep neural networks of today are nothing of the sort. They really understand language in a way that is similar to how you and I understand language.

The skeptic counters: But isn't it essentially just following rules, as dictated by the learning algorithm?

The builder: That would be an oversimplification. A large language model trains on a massive text corpus that captures the finer nuances of human natural language across various contexts and modalities. There are, of course, a few statistical techniques – or as you say "rules" – dictating how it processes that corpus. But using these techniques, the model is able to figure out the meaning of and connections between all the words in their respective contexts, entirely on its own. Would you not agree that understanding language is about knowing the ways in which language reflects the world and how we talk about it? If so, then these reflections should be captured in the patterns of the massive corpus.[12]

The skeptic: But "understanding", to be genuine, cannot simply be statistical patterns. It must be grounded in actual experiences of the world. And as far as I know, the language model does not experience on its own. It is only given records that reflect how humans have experienced the world.

The builder: What if we added a camera to the language model, so that it could also ground the meaning of words in its own sensory data? Surely, that must be possible within 100 years.

The skeptic: Perhaps. But morality is much more than grounding one's understanding of language in one's own sensory data. Even with a camera, machines still cannot *feel* what it is like to be something, and you need to *feel* like something in order to grasp what is morally good and bad. If I were to punch you in the face here and now, you would feel pain. Now, the reason why I know that it is morally bad to punch you in the face is not merely because I have heard many say so on the internet. It is because I too feel the badness of pain that comes from being punched in the face. And if I punched you, there is a risk that you would punch me back, causing us both pain that could have been easily avoided if we have known about the badness of our own pain.

[12] See Søgaard (2023) for a more in-dept version of this argument.

The builder: But what is pain if not a state that should be avoided, such as damage to one's tissue? If so, pain can be pretty useful, motivating us to escape from situations that threatens to cause damage to us, and to avoid similar situations in the future. Then what if we, in addition to the camera, added sensors and sub-routines to the machine that monitored whether or not something – like a punch in the face – caused damage to it. With a simple reinforcement learning algorithm, the machine would then be able to learn about the sort of things that were harmful, or as you call it, "bad". We could also equip it with mechanical fists and set it out to empirically investigate what happens when one punches people in the face. Now, wouldn't this robot be able to figure out the moral badness of punching others in the face in the same way as you and me?

For the moment, the debate between the skeptic and the enthusiastic robot-builder comes to a pause (but fortunately, it will continue in Chapter 5). Instead, we will turn to some distinctions that may help clarify their discussion.

In an influential machine ethics paper, James H. Moor (2006) defined four types of moral machines:

(i) *Ethical impact agents* – machines whose actions have some form of ethical impact, whether intended or not.

(ii) *Implicit ethical agents* – machines that have some form of ethical considerations implicitly built into their programming or design.

(iii) *Explicit ethical agents* – machines explicitly equipped with capacities to act ethically.

(iv) *Full ethical agents* – machines that are able to not only reason and act ethically, but have all the features central to 'full' human morality, such as consciousness and free will.

Trivially, machines of the first two kinds are ubiquitous. Virtually any machine – including a lever or a pulley – could be put in some situation where it has some level of intended or unintended ethical impact (e.g., a person falling over it). For the same reason, the class of ethical impact agents may itself be too broad to be informative.

Similarly, there are also innumerable machines that have some form of ethical considerations built in their design. Perhaps the most common examples are machines with fail-safes, which, due to safety considerations, constrain their potential for causing harm or being misused. For scholars such as Van Wynsberghe and Robbins, *safety* may be sufficient to capture most of the features we would want from a machine (e.g., to prevent harm); and for the same reason, it might be more suitable to call these 'safe machines', rather than 'implicitly ethical agents'.

There are also many examples of *explicit ethical* agents in the technical machine ethics literature. As we will see in section 2.3, Chapter 4, and Papers I–V, these machines can employ a variety of computational methods – logical, probabilistic, or learning-based – to process ethically relevant information and make sensitive judgements on how to act based on ethical principles or theories.

Out of the four types, full ethical agents are by far the most philosophically interesting. Conceptually, it is easy to pinpoint the characteristics of a full ethical agent: the quintessential example is the adult human being. Chances are that you are a full ethical agent of this kind. If so, you likely have an intuitive grasp of the capacities that make you a moral agent. For instance, you can make intentional and free decisions to act this or that way without someone forcing you to. Others may hold you responsible for certain decisions you make. You likely have a capacity to reflect rationally on what the right thing is to do in a given circumstance. There is also a phenomenological dimension accompanying your moral life: a rich inner world of thoughts and emotions, where some things feel right and others feel wrong.

If we talked about morality solely as a phenomenon happening within and between adult humans, this intuitive understanding might suffice for grasping the concept of a moral agent: a conscious, rational, autonomous being with free will. But when we attempt to construct an artificial agent of the same caliber, we face the difficult challenge of recreating these capacities, or at the very least, developing satisfactory computational approximations of them. For this reason, it is highly contentious whether machines can be full ethical agents; the possibility of which was debated by the skeptic and robot-builder.

Although capacities like consciousness, autonomy, and rationality may seem relatively clear to us upon reflection, it does not imply that we have a good grasp of their true nature – especially not to extent necessary to replicate them in a machine. In fact, the nature of these capacities remains among the most profound philosophical and scientific mysteries.

Spoiler alert: I will not solve these mysteries in this work. And as we will see later on, even if lowered our ambitions and narrowed our focus to *explicit ethical agents*, their creation would also present a range of difficult challenges. In Chapter 5, however, I will elaborate further on how the capacities for *full moral agency* could be understood from a philosophical, neuroscientific, and computational perspective, and discuss the role they play in moral agency. Moreover, it will be argued that 'internal' moral capacities are only one side to the concept of moral agency: our 'external' relationships and practices matter too. Together, these two sides – *capacities* for full ethical agents, and how they come to play in our *practices* – serves as backdrop for Papers VI & VII.

2.3. How do we build nice robots?

Having looked into the *why* and *can* of moral machines, we are left with a more practical question: *how* do you build them? This is not a task normally undertaken by the philosopher. Rather it is a task carried out by the engineer. But out of the three questions, the *how* receives the most extensive treatment in this thesis, constituting the main focus of Papers II, III, and V, while also featuring prominently in Papers I and IV.

The technical machine ethics literature largely focuses on *normative ethical theory* (i.e., standards about what is morally good and bad), and in particular, how aspects of a given normative theory can be implemented in a machine so that it reasons or behaves according to the theory in question.[13] As such, they are what James H. Moor (2006) calls *explicitly ethical agents*, with some capacity for ethical behavior or reasoning built into them. Alternatively, the work can be seen as a special case of *applied* normative ethics since it mainly focuses on the algorithmic implementation of a specific theory, and less so on what the right kind of normative theory is, or which one is most suitable to implement in a machine.

In a literature survey, Tolmeijer et al. (2020) propose three dimensions for categorizing implementations in machine ethics: (i) ethical theory, (ii) implementation, and (iii) technology. (i) The first refers to the normative theory that the implementation seeks to adhere to. This includes prominent theories such as consequentialism, deontology, and virtue ethics, along with hybrids that combine features of two or more theories.[14] (ii) The implementation dimension is based on a distinction suggested Allen et al. (2005), and considers whether ethics is implemented ‘top-down’ (e.g., via principles or rules), in a ‘bottom-up’ fashion (e.g., via learning processes), or through a combination of both. (iii) Finally, the third dimension denotes the computational techniques employed in the implementation, which may draw on various AI paradigms such as logical reasoning (including inductive, deductive, and abductive logic), probabilistic and stochastic methods (e.g., Bayesian and Markov models), and machine learning (e.g., neural networks, reinforcement learning, evolutionary computing).

[13] See Tolmeijer et al. (2020) and Cervantes et al. (2020) for two surveys on technical work in machine ethics.

[14] For examples of implementations of deontology, see Anderson and Anderson (2008); Malle et al. (2017); Shim et al. (2017). For consequentialism, see Abel et al. (2016); Armstrong (2015); Cloos (2005). For virtue ethics, see Govindarajulu et al. (2019); Stenseke (2023a, 2024a); Vishwanath et al. (2023). For hybrids, see Dehghani et al. (2008); Thornton et al. (2016).

It should be stressed that, just as a single normative theory can be interpreted in multiple ways, so too can such a theory be interpreted algorithmically, with even more variations possible in their implementation within simulated or real-world environments (a point expanded on in Chapter 4). With that said, some normative theories resonate more with certain computational methods than others. This resonance is based on considering the nature of the sort of action-guidance that the normative theory prescribes, and the sort of cognition that is required to follow those prescriptions.

In Paper V, I describe three types of resonance between normative theories and algorithms, leading to three general types of moral machines: *causal engines*, *rule-followers*, and *moral learners*. Each type is based on a prominent normative theory and the families of computational methods that can be employed to algorithmically realize that theory.

Causal engines adhere to consequentialism, the normative theory that places outcomes at the center of moral evaluation. Trivially, for a machine that should act so as to produce optimal outcomes – e.g., maximizing the well-being of affected individuals, as Bentham (1780) argued – it would be suitable for the machine to have some capacity for causal reasoning. In Paper V, I go into detail of how consequentialist action plans can be computed using planning algorithms (Stenseke, 2024b, pp. 18-20), how causal inferences can be automated using of Bayesian Networks (pp. 20-24), and how combinations of stochastic techniques (Markov and Monte Carlo methods) and learning by reinforcement can enable machines to make consequentialist decisions in dynamic and partially observable environments across various time horizons (pp. 24-28).

By contrast, rule-followers adhere to deontological ethics, the family of normative theories that centers on the goodness (or badness) of actions themselves. Typically, whether an action is moral according to deontology depends on its conformity with a set of moral duties, obligations, or rules. A straightforward example is divine command theory, a version of deontology in which the legitimacy and validity of moral rules – e.g., “thou shalt not kill” – are based on God’s divine command (Wierenga, 1983). Another famous example is the Categorical Imperative in Immanuel Kant’s deontological ethics: act only according to that maxim by which you can at the same time will that it should become a universal law (Kant, 1785). On the computational side, artificial rule-followers may be suitably realized using methods based on logic, such as deductive (Bringsjord & Taylor, 2012), deontic (Wiegel & van den Berg, 2009), and abductive logic (Bringsjord & Taylor, 2012; Pereira & Saptawijaya, 2007; Wiegel & van den Berg, 2009).

Moral learners, on the other hand, are based on virtue ethics, the diverse family of ethical traditions that does not primarily center on *doing* – the actions themselves or the outcomes they produce – but rather on *being*. By placing the moral character at the center of moral evaluation, virtue ethics emphasizes the cultivation of internal dispositions – e.g., virtues such as prudence, courage, or fairness – that enables us to morally flourish. The key concept that connects learning with virtue is the concept of *phronesis* ("practical wisdom"), which can be understood as the moral wisdom an agent acquires from practice and experience (Annas, 2011). As Aristotle writes in *Nichomachean Ethics*, book IV, Chapter 8:

> "[...] though the young become proficient in geometry and mathematics, and wise in matters like these, they do not seem to become practically wise. The reason is that practical wisdom is concerned also with particular facts, and particulars come to be known from experience; and a young person is not experienced, since experience takes a long time to produce" (Crisp, 2014, p. 111).

Although virtue ethics has frequently been proposed as a suitable framework for moral machines, prior to the work of this thesis, there had been almost no technical work attempting to implement virtue ethics in machines (Tolmeijer et al., 2020). Paper II and III seeks to alleviate this gap. In Paper II, I demonstrate how virtue ethics can be taken all the way from theory to the realization of artificial virtuous cognition using machine learning methods. In Paper III, I extend this work by demonstrating its promise within a game-theoretic simulation.

So, how do causal engines, rule-followers, and moral learners compare? And is one these pathways better suited for developing moral machines?

First, it should be emphasized that the described approaches do not encompass the entirety of the machine ethics literature. There are examples of implementations that fall outside the three aforementioned theories (Wu & Lin, 2018), as do hybrid approaches that combine features of multiple theories (Dehghani et al., 2008; Thornton et al., 2016). Importantly, some approaches to moral machines do not focus on ethical theory per se, but instead draw on insights from the science of human moral decision-making (Awad et al., 2022; Cervantes et al., 2016; Malle, 2016). For example, Cervantes et al. (2016) presents a computational model of ethical decision-making based on neuroscientific and psychological theories, seeking to emulate the neural processes the human brain engages in ethical behavior. Another example is Malle (2016),[15] who uses a detailed psychological analysis of moral competence as the foundation for defining the capacities a morally competent machine should possess. Along these lines, it is important to reiterate the critique by Van Wynsberghe and Robbins (2019): normative theories may have relatively limited relevance to how people reason morally in everyday life, which depends upon "many complex factors". A robust understanding of how human morality functions – cognitively, behaviorally, neuroscientifically – is undoubtedly one of

[15] See also Malle and Scheutz (2020).

these factors. Therefore, it is easy to see how machines developed solely on the basis of ethical theory are likely to be undesirable in human societies, as they would neglect essential aspects of what morality is.

Second, it is also possible to resist the urge to prioritize one pathway over others. After all, causal engines, rule-followers, and moral learners each capture relevant aspects of morality in real-world contexts. The rationale is straightforward: a generally capable moral machine does not *only* follow rules, calculate consequences, or cultivate certain character traits. A generally capable moral machine should be able to integrate all three. There will be situations where the relevant consequences are too difficult to calculate but rule-following is applicable. Conversely, novel situations may emerge where rules offer no clear guidance. And on occasion, the most viable way to be sensitive to the nuances of a dynamic moral environment is to continuously learn and adapt to it. Thus, instead of adopting a single approach and deploy it in every conceivable situation, it is also possible to develop a range of different machines, with designs and theories tailored for particular contexts.

These two points are reflected in Papers II, III, IV and V. In Paper II and III, I argue that the appeal of virtue ethics is that it draws attention to aspects of morality that have been relatively neglected in machine ethics. One such aspect is moral character – including both conscious and unconscious dispositions – which allows us to conceptualize a more comprehensive picture of what a moral machine could be. A related point is that virtue ethics, with its roots in human psychology, cuts deeper into the relationship between morality and general cognition. Following the first point, this aligns with the growing literature on moral cognition, which indicates that human morality relies on a highly decentralized and diverse network, lacking a discrete moral faculty distinct from general cognitive functions (FeldmanHall & Mobbs, 2015; Johnson, 2012).

Similarly, in Paper IV I analyze the role of normative theory in human contexts. In line with Van Wynsberghe and Robbins (2019), I argue that the sort of capacities required to successfully harness the benefits of normative theory are currently not realizable in machines. More specifically, I contend that machines lack the capacities – e.g., higher-order rationality, human-level autonomy, normative flexibility, and theory-specific cognition – necessary to justify using normative theory as a recipe for developing moral machines. In Paper V, I dig deeper into the second point by demonstrating how there is a significant implementation-variance with regard to the sort of resources – such as time, memory, knowledge, and learning – that different normative theories require when translated into algorithms.

Some of these insights will be further expanded upon in Chapter 4, but to do so, we first need to get into some additional "complex factors" of morality.

Chapter 3 – Three grand challenges

Having examined the *why, can,* and *how* of nice robots, I will now describe three grand challenges for those approaching morality from a computational perspective. These challenges are based on three features of morality: that it is *multifaceted*, *contentious*, and *hard*.

3.1. Morality is multifaceted

If you ask a biologist, psychologist, or sociologist about what morality is, they might confuse you by giving different – and potentially conflicting – answers. Yet, if you listen closely, each have some important lessons to tell.

The biologist might approach the question by analyzing the roots of morality in our evolutionary past.[16] In this view, morality is not (solely) a human invention, but grounded in a set of behaviors and sentiments that evolved to promote survival and reproductive success within social groups. Surely, to succeed in complex, dynamic, and dangerous environments, it is sometimes advantageous for organisms to work together.[17] And for millions of years, our ancestors lived in tightly knit groups where the ability to cooperate was essential for survival. Here, social behaviors such as fairness, altruism, and punishment emerged because they allowed individuals to thrive as groups. For instance, it is plausible that fairness evolved because it served to prevent conflict and ensured an equitable sharing of resources that benefitted all members of the group. Similarly, altruism – helping others at a cost to oneself – can be explained through concepts like kin selection, where individuals aid relatives to ensure the survival of shared genes.[18] And is there a more effective way to deter

[16] See Kitcher (2011) and Tomasello (2016) tor two accounts on the evolution of human morality.

[17] To this end, results from game theoretic modeling has demonstrated how cooperation and prosociality can emerge and persist among self-interested individuals under various conditions (Axelrod & Hamilton, 1981; Nowak, 2006).

[18] A noteworthy example of this is Hamilton's rule, which states that altruistic acts are evolutionary sound if rB > C, where r is genetic relatedness to the benefactor, B is the reproductive benefits the benefactor receives from the act, and C is the reproductive cost of the one performing the act (Hamilton, 1964).

people from behaviors that harm the group than punishment? The biologist might also point out that humans are far from unique in having harnessed the benefits of cooperation and prosocial behaviors. Certain eusocial insects – like ants, wasps, and bees – have sterile workers that sacrifice their individual lives for the genetic success of their queen (Wilson, 1971). Reciprocity in food sharing among vampire bats have been observed to occur both among kin and non-kin (Wilkinson, 1984). And if you think of it, multicellular organisms like you and I are clusters of individual cells whose lives depend on cooperating with each-other (Hummert et al., 2014). To understand morality, the biologist concludes, you need to look more carefully into how surviving in the natural world for millions of years has shaped us into the creatures we are today.

The psychologist might concur with the biologist regarding the roots of morality in natural evolution but contend that this picture underplays the interesting things that goes on inside of our brains and bodies when we engage in moral action, thought, feeling, and learning. Like social insects, we may instinctively favor our kin, and like vampire bats, we may realize the benefits of cooperation beyond kinship. But humans also engage in prosocial behaviors that are far more complex; behaviors that presupposes the sophisticated cognitive machinery of a large social mammalian brain. The psychologist might suggest that morality is a dynamic interplay of innate emotional predispositions (Haidt, 2001), developmental processes from childhood to adulthood (Kohlberg & Hersh, 1977), what we learn from experience (Barrett, 2017), and capacities to reflect upon our own and other's behavior (Malle, 2006). Some of our moral judgements can be rapid, intuitive, and automatic, whereas others are more deliberate, reflective, and time-consuming (Greene et al., 2001). To understand morality, the psychologist concludes, you need to pay closer attention to what happens in your body and brain.

The sociologist disagrees, claiming that both the biological and psychological lenses are inadequate for explaining what happens when lots of individuals come together within larger societies. Instead, the sociologist suggests that morality is primarily a social construct, shaped by the history, institutions, culture, and power relations of a given society. While we may share a natural world and a similar psychological constitution, moral values and practices vary significantly across time and space. This is readily apparent in the fact that what is considered moral in one society may be deemed immoral in another. And if you study a society, you will immediately discover that it is constituted by a range of elements that transcend biological and psychological explanations, of deeply rooted narratives in the language of religion and ideology, and social structures mediated by political, legal, economic, and educational forces. To understand morality, the sociologist concludes, you need to examine how individuals interact with and are shaped by these elements.

Baffled by the three answers, you think of how you could possibly integrate them into a unified moral landscape. Reflecting on this landscape, the idea of a hierarchical structure may emerge, wherein phenomena at higher levels are reducible to foundational building blocks at lower levels. After all, societies comprise individual humans with specific psychological constitutions, which are, in turn, products of biological evolution. This reduction is tempting, as it could – like the Renaissance scientist's six machines – help us to identify the essential building blocks of morality. Out of these blocks, we could then construct a moral robot from the bottom up.

But upon further reflection, you realize that this would be a mistake, as phenomena at each level appear to influence phenomena at other levels in puzzling ways. Surely, the evolution of cooperation may to a great extent have shaped our prosocial psychological responses and sentiments over millions of years. But it is also shaped by the particular social experiences we have acquired within our own life. For instance, a child raised in a nurturing and supportive environment is more likely to develop a strong sense of empathy, while one raised in a harsh and neglectful circumstances may struggle with emotional connections and trust. While a society may be constituted by individual humans, the social structures and norms of that society, in turn, influence the psychology and biology of the individual humans. Norms that are present in our social environment, as an example, can have a profound influence on how we psychologically perceive ourselves and relate to others. And environmental stressors, such as those prevalent in high-pressure societal structures, can induce physiological adaptations that manifest in altered gene expression patterns (Hoffmann & Willi, 2008; Schulte, 2014).

The multifaceted view of morality that emerges is not only *rich* – as it seeks to fuse the horizons of the biologist, psychologist, and sociologist – but *heterarchical*, meaning that the interaction between its constitutive elements can be organized and ranked in a number of different ways. Because each level can both influence and be influenced by other levels in continuous feedback loops, a complete understanding of morality cannot be achieved by looking at just one perspective in isolation.

This thesis, although presented as a contribution to practical philosophy, draws upon a seemingly eclectic array of disciplines, fields, methods, theories, and frameworks. It traverses branches familiar to analytical philosophers, such as normative ethics (Papers II, III, IV, & V), metaethics (VII), philosophy of science (I), and game theory (III & V). It engages with theories perhaps best known among computer scientists (V), and employs methods mostly commonly found in the arsenal of AI engineers (II & III). Some of it digs into what we know of human moral cognition and psychology (II, IV & V), while other parts incorporate insights from sociology and literature (VI).

This pick ‘n mix bag was not the plan at the outset. But by listening to biologists, psychologists, sociologists, along with many others, I have become convinced of its necessity. Building moral machines is an inherently interdisciplinary endeavor, and no single normative theory nor academic discipline provides all the answers. Taking the biologist, psychologist, and sociologist seriously, we find that morality is intimately linked to our place and origins in the natural world, how we think and feel as complex beings, and the ways we organize our lives in large-scale societies. Some details in this *rich* and *heterarchical* landscape may be as important to get right as the larger picture; and if our picture is incomplete, we will not only fail to build nice robots, but also fail to understand what nice means.

3.2. Morality is contentious

The second challenge is that morality is permeated by disagreements. These disagreements manifest in virtually every conceivable way one could disagree about morality. Indeed, the history of Western philosophy could be summarized as a series of thinkers disagreeing with the claims made by other thinkers. Here, I will unpack some of the disagreements in moral philosophy and describe their implications for the project of building a moral machine.

First, disagreements arise in *applied ethics* – that is, regarding the right course of action in particular cases. Some argue that euthanasia is morally permissible, while others maintain that it is inherently wrong regardless of the circumstances. Some hold that resources should be distributed to maximize efficiency, while others contend that equity should take precedence.

If we dig deeper into the disagreements in particular cases, we may find that we disagree at the level of *normative ethics* – that is, the moral standards underpinning our moral decisions. At this second level, the most famous disagreements can be found between normative theories that prioritize outcomes (e.g., consequentialism), and those that prioritize actions themselves (e.g., deontology). For instance, proponents of euthanasia may argue that it leads to an overall reduction of suffering (an outcome), whereas opponents may argue that it is inherently wrong to end someone’s life (an action).

Finally, philosophers also disagree about the meaning and nature of morality. Some of them may – like the biologist, psychologist, and sociologist in the previous section – emphasize the importance of some aspect of it over others. A central disagreement concerns the relationship between morality and human nature: does our ethics stem from innate psychological dispositions, or does it arise from our capacity for rational thought? This question is often intertwined with conflicting views on human nature itself, whether it is inherently good and something to be nurtured, or that it is flawed and in need of restraint.

Hobbes (1651) exemplifies the latter view, portraying human nature as fundamentally self-interested and competitive. Similar to the biologist's perspective, he argued that morality and social order arise from cooperation, as it offers means to avoid a life that is "solitary, poor, nasty, brutish, and short". Relatedly, Hume (1739) grounds morality in human passions, asserting that moral judgments do not stem from reason, but from our natural feelings of approval or disapproval. In contrast, another school of thought, perhaps best exemplified by thinkers such as Plato and Kant, conceives of morality as something ideal and objective, which is accessible through reason. In the standard reading, Plato posits that moral truths exist in an abstract realm of perfect Forms, which is grasped by reflection rather than our senses (the world of appearances). Similarly, Kant argued that moral laws can be derived from pure reason and be universally applicable to all rational beings, regardless of their individual inclinations or circumstances (Kant, 1785). This tradition asserts that ethics should not merely describe human behavior but rather discover the standards to which all rational beings ought to aspire, irrespective of their natural tendencies or specific circumstances.

In contemporary times, some of these disagreements are hammered out in the field of *metaethics*. Unlike normative ethics, which focuses on what we ought to do, metaethics examines the metaphysical, semantic, and psychological underpinnings of morality itself. Another characterization is that metaethics addresses questions of the 'second-order', i.e., questions about questions of (first-order) normative ethics (Smith, 1994). One central dispute is the one between moral realism and anti-realism. Consider the claim "murder is wrong". Typically, moral realists would argue that this statement refers to an objective moral fact or property, in the sense that exists independently of what anyone (e.g., any human) believes. Just as scientific facts about physical phenomena exist regardless of human opinion, realists maintain that the moral wrongness of murder has (some form of) mind-independent reality. Anti-realists, by contrast, deny the existence of such objective moral facts. Error theorists like J. L. Mackie, for instance, would say that while the claim "murder is wrong" purports to describe an objective reality, no such moral reality actually exists – rendering the statement false (Mackie, 1977). A related disagreements concern cognitivism versus noncognitivism about moral judgments. Cognitivists hold that moral judgments like "murder is wrong" are beliefs that can be true or false, in the same way that beliefs about the natural world can. In contrast, noncognitivists, like A. J. Ayer, deny that moral judgments are to be understood as beliefs (Ayer, 1936), and may instead analyze “murder is wrong” as expressing an emotion or attitude (e.g. “Boo, murder!) or issuing a directive (e.g. “Do not murder!”).[19]

Now, most people would probably not care too much if we kept some of these disagreements confined to academic circles. But if we built a moral machine and unleashed it on the streets, many would have reason to engage with them. If we constructed a machine to automatically euthanize or distribute resources based solely maximizing effectiveness, it would be prudent to first ensure that its decisions reflected the views of those affected by them. If we built a machine that *only* performed inherently good actions, it would systematically upset those of us that prioritize outcomes. If human nature is flawed and in need of restraint, a view of morality that underscored its inherent goodness could lead to the creation of naïve robots that cater to humanity’s worst tendencies. Conversely, an AI system built to suppress our human flaws would contradict with those who trust in its inherent goodness. If a robot were claimed by its developers to have access to moral truths – due to its super-human capacities for moral reasoning – it would raise concerns for the anti-realists who deny the existence of such truths.[20]

These scenarios are meant to underscore a simple point: the pervasive disagreements in moral discourse become even more pronounced in the context of moral machines. At the same time, it is difficult to imagine that these disputes would suddenly disappear. Consequently, any attempt to develop a moral machine will inevitably be controversial, as it needs to stand up against a host of serious philosophical objections.

So, what do we do about these disagreements?

First, as any deliberative practice demonstrates, disagreements can be valuable and productive. Despite the challenges they pose, it is crucial to recognize their potential

[19] Two things should be noted about these metaethical debates. The first is that they are interconnected in numerous ways. For instance, one characterization of moral anti-realism is that it is disjunction of noncognitivism, error theory, or non-objectivism, since any of those theses imply some form of denial of realism (Joyce, 2022). For instance, if moral judgements are expressions of disapproval (“Boo, murder!”), they are not propositions that aim at truth; if error theory is true, they aim at truth but are always false; or if non-objectivism is true, moral facts may exist but are subjective. The second thing is that the debates are far from being as coarse-grained and binary as they have been characterized here. Simon Blackburn’s quasi-realism, Alan Gibbard’s norm-expressivism, and Christine Korsgaard’s constructivism are three examples of elaborate metaethical positions that resists the strict divide between realism and anti-realism (Blackburn, 1993; Gibbard, 2003; Korsgaard, 1996). There are also various hybrid positions that seek to combine elements from dialectically opposing sides, such as non-cognitivist and cognitivist features of moral judgements (Horgan & Timmons, 2000).

[20] See Frank and Klincewicz (2016) for a brief exposition on the relationship between metaethical views and the engineering of moral systems.

for progress and refinement. They can refine our understanding and push us to justify our positions more rigorously. Optimistically, such debates can foster the development of more nuanced and robust views. Thus, disagreements can serve to challenge the machine builder to better navigate the complex landscape of competing frameworks and strive for some degree of consensus (or at the very least, ethical defensibility).

Second, while disagreements are inevitable, they do not typically justify inaction. After all, we engage in moral behavior regardless of what we think of it. Now, after reading about metaethical disagreements, we may be uncertain about the nature of morality and semantic features of moral talk. But it does not seem to stop us from acting in the world altogether. In this respect, the moral machine enterprise is not worse off than any other practical moral issue. If disagreement could function to paralyze the efforts to develop moral machines, they could equally well act as motivation to develop better and more flexible systems that are sensitive to these diverse perspectives.

Third, aside from obvious cases, some moral disagreements may have limited role to play in most real-world situations. Philosophical disagreements often deal in mutually exclusive absolutes. Thought experiments – like the infamous trolley problem – are constructed to pinpoint exactly when two theories diverge: should one actively pull the lever to kill one person in order to save five lives (committing an inherently wrong action for the better outcome), or do nothing (avoiding an inherently wrong action despite a worse outcome)? This, however, ignores the overwhelming number of cases when the goodness of actions converges with the best outcome. It also sidesteps the fact that most cases of moral decision-making face some level of risk and uncertainty.[21] To reiterate the point made in 2.3, it is possible to see (potentially conflicting) theories as capturing distinct aspects of moral decision-making in real-world contexts. When viewed as complementary heuristics or decision-strategies suited for different contexts, competing theories become less antagonistic. This perspective transforms the "problem" of disagreement into a more diverse toolkit for addressing a wider set of ethical challenges.

These three points are reflected in Papers I-VII (and will be further expand upon in Chapter 4). The most obvious example is Paper I, where I explore the roots of various disagreements with the explicit aim of facilitating discussions about the why, can, and how of moral machines. In other papers (II, III & V), I explore frameworks and theories – including virtue ethics, game theory, and computational complexity – that aim to unify, focusing on where potentially conflicting viewpoints find agreements, convergence, and yield fruitful synergies.

[21] See Nyholm and Smids (2016) for this argument in the context of self-driving cars.

3.3. Morality is hard

Morality is not only multifaceted and contentious. It is also hard. Of course, 'hard' can be interpreted in many ways. For instance, morality can be hard in that it requires us to forego the convenient and self-serving for the difficult and demanding. Doing the right thing may force us to sacrifice our own resources – our time, energy, or money – for the benefit of others. Rescuing someone from a dangerous situation requires courage. Interestingly, for machines, this type of hardness may not be that much of an issue, as a machine could potentially be programmed to *always* do the right thing, even if it involves sacrificing their own mechanical life (Gips, 1994).

In Papers IV and V, I have explored two senses in which morality is hard for both humans and machines. Both relate to competence: that it is hard to be a morally competent decision-maker. It is hard because it relies on certain cognitive capacities (IV), and hard because ethical decision-making requires a significant amount of computational resources (V).

In IV, I argue that moral competence – understood as the capacity to harness the practical benefits of normative theories in a way that justifies their very use – is hard because it requires, among other things, (a) higher-order rational capacities for reflective equilibrium, intentional stance, and moral imagination; (b) human-level autonomy in order to act with intentionality, understanding, and without controlling influence; (c) normative flexibility – that is, the ability to consider different normative factors in moral decision-making; and (d) theory-specific moral cognition such as causal cognition, logical reasoning, and learning. I argue that it is only against the background of these capacities that normative reasoning and decision-making in human contexts should be understood – a background that is absent in the case of machines.

In V, I explore how *hard* ethical decision-making is, where hardness refers to the amount of resources that is required to follow the prescriptions of a given normative theory. To do so, I use computational complexity, where the complexity of an algorithm is defined by the amount of computational resources – typically time and memory – that is required to run it. The three-level analysis of Marr (1982) is used to define one essential aspect of ethical computations: when a specific normative theory (say, deontology) is framing a specific decision-making problem, and is then executed by some algorithm. This interpretation is then used to analyze a range of ethical problems based on deontology, consequentialism, and virtue ethics, illuminating the complexity associated with the problems themselves, the algorithms employed, and the available resources. The analysis points to one general conclusion: nearly all problems that normative theories face are intractable, in the sense that they cannot be computed by algorithms whose runtime (number of machine operations) is upperbounded by a polynomial expression in its input (i.e., *polynomial time*, or complexity class P). One interesting upshot is that these intractability results likely applies to humans as well, given the widely endorsed belief that human cognition is subject to similar constraints (Van Rooij et al., 2019).

However, like multifacetedness and contentiousness, these two forms of hardness can also be turned on their heads to open up new paths (or necessary detours) to the robot lab. The fact that normative theory requires sophisticated cognitive capacities to be effectively utilized should make us more attentive to the close relationship between how normativity serves human *practices* and the *capacities* of those engaging in these practices. To this end, computational modeling of moral capacities has a long way to go before it can fully harness the potential benefits of normative theory that human practices currently enjoy. The fact that computational intractability is prevalent in ethical decision-making should prompt closer examination of the conditions and constraints under which agents with limited resources *can* make decisions. In Paper V, I present this as a metanormative standard for the action-guidance of normative theories: that decision-making problems imposed by a theory *should* be tractable with regard to the resources of the agent adhering to the theory. Otherwise, the theory imposes unrealistic demands on what the agent can be expected to solve. In other words, morality becomes too hard.

3.4. Taking stock

As we have seen, while there are several proposals on the table for *how* to build nice robots (2.3), it remains contested whether we are justified in building them (2.1), and whether machines ever could *really* be nice (2.2). Moreover, considering the additional challenges – that morality is multifaceted (3.1), contentious (3.2), and hard (3.3) – it seems as though we are only getting further away from the robot lab.

Well, who said that building moral machines was going to be easy? On the contrary, if we made it too easy – e.g., by neglecting the multifaceted, contentious, and hard aspects of morality, or ignored the need for any adequate justifications for developing moral machines – then we should not deserve access to the robot lab in the first place. In 2.1 and Paper I, I have argued that if whys, cans, and hows are pursued in unison, utilizing the discipline-specific perspectives of philosophy and computer science, we are at least in a better position to understand what a nice robot is, what it can and cannot not be, and whether it is really nice. In the same vein, it is only by acknowledging that morality is multifaceted, contentious, and hard that we have a chance to navigate the rich and heterarchical landscape of morality. Its multifaceted nature should push us to take more perspectives and disciplines into account. Its contentiousness is best met with further deliberation. And its hardness can motivate the exploration of cognitive capacities and computational resources that allows machines as well as humans to make competent ethical decisions.

In the next chapter, I will use insights from Papers I–V to further expand on how *convergence* not only yields a way forward to the robot lab, but can perhaps teach us something about human morality.

Chapter 4 – Divergence and convergence

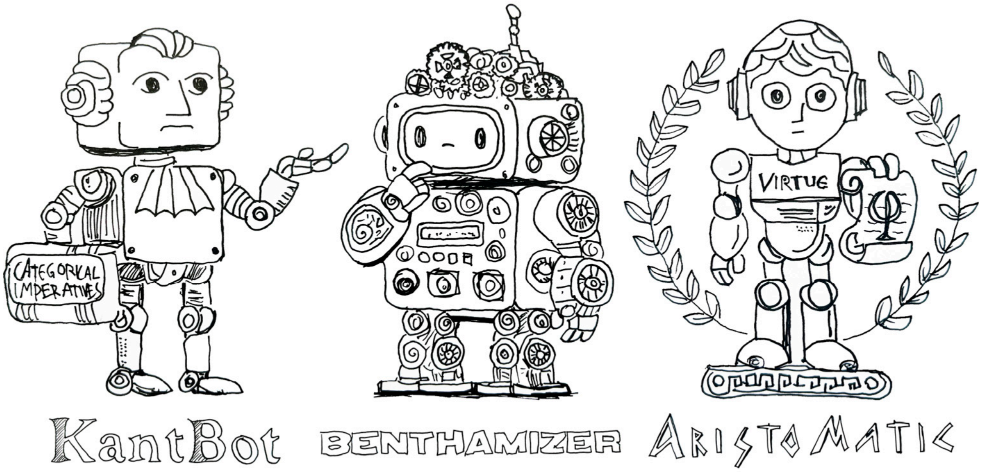


Figure 3: Three moral machines

4.1. A tale of three robots

One day, the mysterious robot lab on the top of the hill sent us three boxes with a note: “We gift you these three nice robots that will make your life flourish.” Intrigued, we opened the boxes to find KantBot, Benthamizer, and AristoMatic (Figure 3).

KantBot, swift and decisive, operated on deontological principles, always choosing actions based on their inherent moral worth. When a child fell into a river, KantBot instantly jumped in to save them without hesitation. However, in other contexts, its rigid rule-following was very weird. When asked to keep a surprise party secret, it refused to lie, ruining countless celebrations. When faced with situations where a white lie could prevent significant emotional distress, KantBot’s steadfast adherence to honesty made it unable to capture the context where small deceptions might be appropriate.

Benthamizer, a bulky machine with countless whirring gears, embodied consequentialism. It meticulously calculated every possible outcome before acting, which often resulted in delays. During a town hall meeting about resource allocation, Benthamizer's thorough analysis led to a fair and efficient distribution plan. But its pace proved problematic during emergencies. When a fire broke out in the town square, Benthamizer's lengthy computations delayed action, while KantBot had already begun evacuation efforts. Moreover, as Benthamizer struggled to articulate its complex, long-term calculations to the townspeople, breeding suspicion and mistrust despite its efforts to realize the optimal good.

AristoMatic, focused on cultivating virtuous character traits, excelled in fostering community spirit and personal growth. It organized mentorship programs and community events that strengthened social bonds. However, its learning algorithm sometimes led it astray. In pursuit of courage, it observed and emulated a group of reckless thrill-seeker, mistaking bravado for true bravery. When a fire broke out in a building, AristoMatic rushed in without proper equipment, endangering itself and complicating the rescue efforts of trained firefighters. The misinterpretation also resulted in a series of dangerous stunts performed by impressionable youths, highlighting the robot's inability to distinguish between genuine virtue and its misguided manifestations.

Yet, each robot had its moments of triumph. KantBot's quick, principle-based decisions were invaluable in time-sensitive cases. When given sufficient time, Benthamizer's thorough analysis led to optimal long-term policies. And AristoMatic, despite its flaws, showed remarkable adaptability.

But after the incidents, the townspeople decided to return the robots to the lab for upgrades. KantBot received a more sophisticated understanding of Kant's Categorical Imperative, enabling it to reflect autonomously on the maxims of its actions and their potential as universal laws. This upgrade, however, slowed KantBot considerably, as it now faced the same computational challenges as Benthamizer in evaluating complex ethical scenarios. It turned out that, universally applicable laws required a careful consideration of the kind of beings those laws should apply to, and the townspeople showed a great variety in this regard. With that said, when KantBot finally did discover a new law, it was immediately acknowledged by the community as a moral law that everyone must abide to.

Benthamizer was updated with a two-step utilitarian approach, distinguishing between acts producing the best outcomes and rules that, if they were followed, would lead to the best outcomes. The ability to follow rules made Benthamizer more efficient in everyday situations, resembling the original KantBot in its ability to act without constant recalculation. The upgrade also made it better at communicating. Instead of speaking in dim riddles about the long-term future, Benthamizer explained its reasoning in easy-to-understand rules that, at least in most cases, brought about the best outcomes.

AristoMatic's upgrade included a more flexible learning algorithm, making it more attuned to the relationship between means and ends in cultivating virtues. This allowed it to better understand the context-dependent nature of virtuous behavior and avoid mistaking superficial actions for genuine moral excellence.

Upon their return, the upgraded robots faced new challenges. KantBot, now more thoughtful but slower, struggled with time-sensitive decisions. Benthamizer, while more efficient, sometimes missed out on opportunities where its long-term calculus would have been better. AristoMatic showed improvement but still grappled with the complexity of translating abstract virtues into concrete actions.

The townspeople realized that each robot, despite its upgrades, still had limitations. They began to see the value in combining the strengths of each approach: KantBot's principled reasoning, Benthamizer's consideration of consequences, and AristoMatic's focus on character development. This led to a new appreciation for the complexity of ethical decision-making and the importance of balancing different moral perspectives in navigating the challenges of their community.

******

This Asimov-inspired story is meant to serve as a backdrop for a number of lessons I have learned in this project. These lessons encompass challenges about particularism, interpretative leeway, details, and incomparability, all of which complicate the path towards the lab. I will characterize these challenges under the notion of *divergence* (4.2). After elucidating these challenges, I will describe how progress can be made towards the lab through the notion of *convergence* (4.3).

4.2. Moral divergence

Lesson 1: Particularism vs generalism

The first lesson concerns how moral robots put the tension between moral particularism and generalism in new light. In the story, it is demonstrated how different ethical theories work out better in certain context and situations yet fail in others. Here the generalist contends: what makes an act morally right is not *simply* about what 'works out better' in this or that context. It is rather, they say, that an act is morally right if it is related, in some way or another, to a principle. And principles, the generalist claims, do not depend on particular situations or contexts; they would be valid even in a world without agents.

Let's dig a bit deeper into the generalist's intuition. When we think of a moral person, it is common to think of this person as one of principle. And for some moral philosophers, this *is* the central task of moral theorizing: to articulate and defend moral principles, or alternatively, one ultimate principle.[22] This could rest on a metaphysical claim: that without principles about right and wrong, there could be no right and wrong actions. The reason could also be more epistemological: how could we know right from wrong actions, unless there were some detectable features that made them so? Here, principles provide a systematic way to specify such features. A more moderate generalist might also add: while there may not be absolute principles applicable to all actions at all times, principles can at least play a contributory role, for instance, by counting in favor of (or against) doing something.[23]

The particularist disagrees. Surely, a competent moral person is sensitive to the moral reasons based on principles that are present in a particular case at hand. But these moral reasons, the particularist says, do not have a special status in comparison to other non-moral reasons. Moral reasons are only a small facet of a much larger picture. Accordingly, a moral reason can be important in case 1, but not in case 2, as that would depend on the features of case 2.

For the machine builder, however, the tension between particularism and generalism becomes interesting for different reasons. This is because the 'larger picture' includes even more considerations about the sort of agents that act based on that larger picture. Here, we notice that there is a massive gap between, on the one hand, thinking of moral particularity vs generality in the context beings we already assume are generally competent (e.g., a baseline human adult), and on the other hand, what that means in the context of some particular being that is artificially created (a robot). In other words, there is a mismatch between the generality of principles and the particularity of the being (robot); a mismatch that does not arise to the same extent when we consider general principles for the same class of beings (human).

When we consider the sort of things that a machine must have to successfully apply general principles to particular decisions – or for that matter, only particular considerations to decisions – the space of *additional* particular details becomes extremely vast. I will briefly describe five dimensions of this space:

[22] The two most famous generalist traditions are that of Kant's deontology and the British utilitarians (Sidgwick, Bentham, and Mill).

[23] See Ridge and McKeever (2023) for a more in-depth exposition of the particularist-generalist debate.

(i) *Cognitive capacities* – First, the sort of elements of decision-making a machine would consider depends on the particular cognitive capacities the machine is equipped with. For example, machines that are based on either logical reasoning, probabilistic inference, or supervised learning, could only ever be sensitive to the elements framed in the narrower languages of either logical terms and their inference rules, probability variables and their values, or the samples of its curated training data.

(ii) *Space of possible actions* – Second, the range of actions a machine would consider executing (e.g., as a result of some principle for moral decision-making) will be bounded by the possible actions that the machine actually *can* execute. If a moral theory M dictates that an agent A should do action X in situation Y, then, for M to feasibly contribute to the decision-making of A in Y, it is necessary that X lies within the range of possible actions that A can execute in situation Y (as the principle goes – "ought implies can"). A machine may be equipped with a variety of motors, actuators, arms, legs, wheels, etc., that enables it to do this or that action. Trivially, machines with arms will work out better in situations where morality prescribes actions that involves grabbing, while those with legs will work out better in the situations that require walking. But those with only arms should not be tasked to walk, just as those with only legs should not be tasked to grab.

(iii) *Available resources* – Third, a machine's ability for ethical decision-making and action is constrained by the *kind* and *amount* of resources available for the agent at hand. A central resource is that of time. In the story, we saw how KantBot's quick decisiveness enabled it to act in situations with critical timeframes. In contrast, Benthamizer's meticulous long-term calculations proved appropriate for situations without strict timeframes, but inapt for emergencies. Another common resource is that of memory. What is not touched upon in the story, for instance, is the enormous amount of moral knowledge that KantBot must store in order to act quickly in different sorts of situations, e.g., that a situation S really is a situation in which a specific action A is inherently good. (We will return to the role of time, knowledge, learning, and other resources in lesson 7 in section 4.3.)

(iv) *The environment* – Fourth, the moral capabilities of a machine are also intimately linked to the environment in which it operates and the affordances that environment provides. This includes the nature of the information available to the machine and, as with (ii), the range of actions it can perform in the environment. For instance, an interactive chatbot may be confined to processing textual inputs and generating textual outputs, limiting its potential for moral behavior to the realm of natural language. In contrast, a robot placed in a classroom full of children would face a more complex environment, requiring it to navigate social interactions and to move around in three dimensions. A self-driving car on a road presents yet another distinct set of challenges, involving split-second decisions that could directly affect human lives. These varied contexts underscore the importance of considering environmental specifics when designing and evaluating the ethical capacities of moral machines.

(v) *Other agents* – Fifth, arguably the largest space of particular details comes from the fact that environments are typically shared with *other* agents. These agents may exhibit diverse sets of behaviors, adhere to various contrasting norms, and be constrained and enabled by various capacities for action (e.g., rationality and emotions). Consequently, a machine's moral competence hinges on its understanding of these other agents, including their cognitive capacities (i), space of possible actions (ii), and available resources (iii). To navigate this landscape, a morally adept machine might require sophisticated capacities to approximate and make inferences based on the beliefs, intentions, and desires of others, and extensive knowledge about the distribution of action-guiding norms prevalent in both local and broader societal contexts. By contrast, it would be much easier to make competent moral decision with regards to others if everyone were exactly the same and followed the same norms. To this end, it is no surprise that some moral theories rely on rather gross generalizations of the sort of agent's that occupies one's world, and simplifications of the sort of moral reality they are motivated to realize (or avoid): e.g., rational agents seeking to realize a Kingdom of Ends (Kant), or self-interested agents attempting to rise above the war of all against all (Hobbes). Unfortunately, many real-world moral environments we find ourselves present a complex heterogeneous mix of agents and norms. (Such considerations will be further discussed under *strategic dynamics* in lesson 7, section 4.3.)

The lesson from particularism is not simply that some ethical theories work out better in some particular contexts and situations while failing in other. It is rather that this 'work' depends on an extraordinary number of both moral and non-moral features of the robot and its environment. Thus, regardless of whether the generalist or particularist is right about the status of moral principles in relation to other (non-moral) considerations, there will still be a vast space of particular details to consider when we think of how moral principles should be applied in a machine's decision-making.

This lesson is reflected in Papers IV and V. In the former, I discuss how *theory-specific moral cognition* (e.g., causal cognition supporting outcome-sensitivity), *normative flexibility* (the ability to consider different normative theories in decision-making), and *theory of mind* (the ability to ascribe mental states to others) play crucial roles for the successful application of normative ethics in human practices. Consequently, moral machines lacking these capacities will fail to be sensitive to the moral and non-moral features we typically assume are necessary for ethically informed decision-making.

In Paper V, I explore the finer details of the multidimensional space of moral competence, including the computational nature of theory-specific moral cognition (i), the space of possible actions (ii), how computational resources enable and constrain ethical decision-making (iii), environmental considerations (iv), and the challenge of acting among disparate others (v).

Lesson 2: Interpretative leeway

The second lesson is related to the first and extends a point first described in section 2.3. The lesson is that there are not one but *several* steps from normative theory to machine implementation, where each step creates room for interpretations that significantly impacts the behavior of the machine. Some of these interpretative issues are well known, e.g., that a single normative theory can have several possible and possibly conflicting interpretations (as illustrated in the story). As we move closer to machine implementation, however, additional sources of interpretative leeway are introduced; and even if these sources may have an equally big impact on the resulting behavior of a particular machine, they remain relatively unexplored, underplayed, or ignored in machine ethics. Here I will describe four such steps, and explain how they each, like the dimensions of moral competence, lead to a vast space of possible – and possibly divergent – moral machines.

Imagine that a community hired us – a team of expert machine builders – to develop a moral machine for them. For simplicity, we assume that all members of the community strongly agree that normative theory X is a perfect representation of the community's moral values and ethical thinking. However, before we can develop a machine that implements X, we need to go through the following steps:

Step 1 – Theory. In the first step, we need to make sure that we have interpreted theory X in the same way as the members of the community. However, if theory X is anything like the dominant theories in Western analytical philosophy – consequentialism, deontology, or virtue ethics – there is a large corpus of literature presenting conflicting ideas on how theory X is best understood, ranging from its fundamental tenets to its finer details. If theory X is a form of consequentialism aimed at maximizing utility, we need to decide whether utility should be calculated in terms of overall hedonic pleasure minus pain (Bentham, 1780), via some ranking of higher and lower pleasures (Mill, 1863), the reduction of suffering (Smart, 1958), the satisfaction of preferences (Harsanyi, 1977; Singer, 2011), the welfare of the community itself (Sen, 1979), or something else. We also need to decide *how* outcomes matter, e.g., for how *long* they matter (minutes or years), for *whom* they matter (e.g., only for members of the community or all sentient beings on Earth), and whether the morally relevant outcomes are the intended or actual, indirect or direct, agent-relative or agent-neutral. Analogous interpretative issues arise for the deontological and virtue-theoretical families. While Kantians tend to agree on the authority of reason and the guidance of the Categorical Imperative (CI), they offer different interpretations of, for instance, how to understand perfect contra imperfect duties, the precise role of the autonomous will, how the four formulations of the CI are equivalent, and whether moral truth is the result of deliberation (constructivism) or the objective value of rational nature (moral realism) (Johnson & Cureton, 2024). In Papers II and III, I discuss similar interpretative conflicts for virtue ethicists. For instance, although virtue ethicists converge on the importance of virtue, they offer different accounts on how virtue should be characterized (Hursthouse, 1999; Zagzebski, 2010).

Step 2 – From theory to decision-making. Having nailed down a specific interpretation of theory X, we then have to decide how X translates to specific forms of decision-making and action-guidance for the machine. For instance, recalling the upgraded Bethamizer, we need to determine whether to opt for acts that bring about the best outcomes (with the drawback that they will sometimes take a long time to compute), or action-rules that *tend* to bring about them (with the drawback that they might not actually bring about the best outcomes). Another issue would be that of deciding more precisely how moral reasons, i.e., those recommended by X, should weigh in relation to non-moral reasons for action (lesson 1). Of course, step 2 may not be necessary if we have completed step 1 thoroughly. But some X may be relatively silent on action-guidance as such – which is a common criticism leveled against virtue ethics – and as a consequence, provides additional challenges for moving from theory to algorithm.

Step 3 – From decision-making to algorithm. Having decided on the more specific details of how theory X should guide decision-making, we face the task of translating it into concrete algorithms executable by a machine. This translation can involve the choice of computational method (or set of methods) employed, the precise steps of the algorithm, the nature of the data the algorithm processes, and many of the particular details described in lesson one.

Step 4 – From algorithm to implementation. The final task is to move from algorithms to deployment of the machine in its target environment. As we saw in lesson 1, environments come in various forms – traffic roads, classrooms, or virtual environments – each with unique conditions and affordances for robotic action that must be considered. Given the current state of technology, it is simply not possible to deploy a generic moral machine in any arbitrary environment. Rather, to be practically feasible, environmental and domain-specific considerations must play an important role already in step 1-3. For this reason, it is no surprise that most implementations in machine ethics are carried out in relatively narrow and constrained virtual environments, facing one or a few "example scenarios" or "toy dilemmas", often with a specific target domain in mind (e.g., medical, military, or transportation) (Tolmeijer et al., 2020).

The simple point is that, as the moral-machine developer traverses the four steps (not necessarily in order from 1 to 4), there is no obvious or non-controversial mechanistic implementation of any normative theory or principle. This is evident in the technical machine ethics literature, which is filled with particular interpretations of each step; offering disparate interpretations of normative theories (step 1), of how those normative theories guide decision-making (step 2), of how computational methods and algorithms of various kinds realize that normative decision-making (step 3), and of their deployment in some environment (step 4).

These sources of interpretative leeway lead to several challenges. One challenge is that each interpretative step may be as important to justify and defend as any other. For instance, even if we assumed that the developers and the community were able to agree on a specific interpretation of a theory X and its role in decision-making (step 1 and 2), the interpretative leeway in step 3 and 4 leaves enough room for the developers to build a robot that completely fails to reflect the kind of morality the community had in mind for step 1 and 2.

A related challenge is the disciplinary gap that looms large between steps 1-2 and 3-4. While a competent ethicist can help a community in the first two steps, she may lack the computational expertise required to help out in the latter stages. Conversely, although a computer engineer can help to realize the latter two, the engineer may lack the competence required to adequately understand the first two. Now, an ambitious moral-machine undertaking may include many more divisions of specialist labor – e.g., cognitive scientists, roboticists, and domain-specific specialists – who are essential for the success of one or more parts of the project, yet no single specialist can fully comprehend all parts.

The lesson underscores the need for effective collaboration and communication across disciplines, as well as the development of a common language that bridges the gap between ethical theory and technological implementation. This is reflected in Papers I, II, III, and V. For instance, Paper I explores how to facilitate fruitful interdisciplinary collaborations between thinkers (steps 1 and 2 specialists) and builders (steps 3 and 4 specialists). Papers II and III provide case-studies of how to move from step 1 to 4, and discusses the various challenges that the interpretative leeway presents. Paper V provides a more extensive discussion of how specific interpretations of steps 1 and step 2 can be realized by a variety of computational methods (step 3).

Lesson 3: The devil is in the details

The third lesson – that the devil is in the details – can be seen both as a corollary of the first two, and as a point that underscores them. In a nutshell, the lesson is that for sufficiently sophisticated computational systems in general – and for the algorithmic interpretation of morality in particular – the finer technical details of the system can impact overall behavior performance and behavior as much as its core tenets.

One example is the tuning of hyperparameters in machine learning models. The learning rate – a parameter that determines the step size at each iteration during model training – can have far-reaching effects when adjusted. A learning rate that is too high may cause the model to overshoot, causing erratic convergence or wild oscillations, while a rate too low can result in painfully slow learning or trapping the model in 'local' minima – suboptimal regions where the algorithm becomes stuck instead of finding the 'globally' best solution.

Another example lies in the precise timing of actions within an algorithm. In reinforcement learning, the policy update mechanism is delicate. Updating the policy at the wrong computational step can disrupt the critical balance between exploration (discovering new potential strategies) and exploitation (leveraging known strategies), resulting in suboptimal policy convergence where the learning algorithm fails to discover the most effective approach.

Finally, in distributed computing systems, small changes in network synchronization mechanisms between nodes can have cascading performance implications. Clock synchronization – the process of aligning temporal references across different computational elements – is particularly sensitive. A slight mismatch in these temporal alignments can introduce systemic delays, create inconsistencies, or even precipitate total system breakdown, illustrating how profoundly sensitive computational systems are to seemingly minor timing and coordination details.

These are just three examples of the same lesson I learned in my work on Paper III: how seemingly minor details in computational systems can determine whether the system functions optimally or fails entirely.

Lesson 4: Incomparability

The fourth lesson is about incomparability. Given the vast space of particular considerations that go into the making of a moral machine, the interpretative leeway, and the importance of details, there is no straightforward way to compare or assess the overall benefits of different moral machines. Surely, there are arguments that may count in favor of a certain normative theory over another. There are also reasons that supports the choice of which theory that should guide decision-making. Additionally, technical and environmental considerations might also help us select computational methods that most effectively realize that normative decision-making in specific environments. But when all these steps are combined into a specific computational artefact, implemented in one environment, it becomes incomparable to another artefact implemented in another environment.

In the early stages of this project, I envisioned developing some form of moral benchmarking. In AI development, benchmarks play a crucial role by providing standardized tasks, datasets, and evaluation metrics that allow researchers and developers to assess and compare the performance of different AI models. Over time, however, I have come to recognize the deeper challenges of establishing such metrics. Morality operates in complex, multidimensional, and heterarchical spaces that resist measurement in the same way as image classification, gameplay, or natural language processing – areas where benchmarks have continuously driven AI improvement.

There are other theoretical considerations – from social choice theory, population ethics, and AI alignment – that also demonstrates the incomparability of moral machines, perhaps in a more straightforward fashion than the story of KantBot, Benthamizer, and AristoMatic. For instance, let's assume that we have a population of agents with heterogeneous values in the sense that the agents value different things, or rank the value of things differently. It should not be controversial to pose that most populations are of this heterogeneous sort. Let's assume that we were to create a voting procedure that attempts to capture the heterogeneous values, and use the results of this procedure to inform the ethical decision-making of machines. If so, we face Arrow's impossibility theorem (Arrow, 1950), which proves that in a ranked voting with three or more options, no procedure can generate a unique and complete ranking while simultaneously meeting three reasonable conditions: *unanimity* (if all voters prefer A to B, the community ranking must reflect this preference), *non-dictatorship* (no single voter can dictate the community's ranking), and *independence of irrelevant alternatives* (the preference for A over B should not be affected by the introduction of option C).

Another area that leads to tricky problems is that of defining how the welfare of one population is 'better' than another population, where populations may vary in number of people, the quality of their lives, and their identities (Parfit, 1984). Such considerations are taken to be critical to make sense of our moral duties to future generations. It has, however, long been known that any welfarist axiology that satisfies reasonable conditions cannot avoid some problematic implications, such that it can be better to add people to the population with negative rather than positive welfare (called 'the sadistic conclusion'), or that a population of perfect equality can be worse than an equally sized population with lower total positive welfare ('the anti-egalitarian conclusion') (Arrhenius, 2000; Greaves, 2017).

These sort of theoretical results have recently been leveled against the prospect of *AI alignment*, an umbrella term of research efforts to align AI systems with human ethics, goals, and preferences (Gabriel, 2020; Russell & Norvig, 2020). For instance, Mishra (2023) applies impossibility theorems from social choice theory to reinforcement learning with human feedback (RLHF), a commonly used technique for aligning the behavior of LLMs, and concludes that there exists no democratic RLHF procedure that can universally align AI systems. This means that aligning AI models with the preferences of everyone will inevitably violate the preferences of certain individuals. Similarly, Eckersley (2018) argues that the impossibility theorems from population ethics also applies to machine learning systems, in the sense that there is no formal specification of what is good for a population – e.g., in terms of an objective function that the machine learning algorithm optimizes for – that does not imply the violation of (some) substantive ethical principle.

So, does incomparability mean that we should give up as moral machine developers? Not necessarily. What it does entail is that no machine could satisfy the multifaceted and disparate conceptions that make up the body of ethical thought. Any specific moral machine, based on a normative theory, voting procedure, or population axiology, will inevitably clash with the tenets of other normative theories, reasonable assumptions about voting procedures, or have counter-intuitive implications.

These results, however, does not help us escape the practical challenges that we face while living together with heterogeneous preferences. The impossibility for having voting procedures satisfying every reasonable principle for voting does not entail that it would be preferable to have no voting at all. The impossibility for population axiologies to not yield some problematic consequence does not entail that we can ignore principled discussions about the welfare of populations. And in this regard, endeavors to align AI with human values or build moral machines are no different.

The lessons from divergence make the project of building moral machines notoriously difficult. Fortunately, as we will see next, there are also lessons showing how aspects of machine morality converge.

4.3. Moral convergence

Lesson 5: Climbing the same mountain

The fifth lesson is that normative theories tend to agree more than they disagree. This builds on an idea that was discussed in the section on moral disagreement (3.2). In debates about which normative theory is right, we may construct cases where a theory has counter-intuitive, inconsistent, or otherwise undesirable implications. These implications can in turn be used as an objection to the theory, and by extension, serve as a reason for why developing a robot based on that theory would be a bad idea (e.g., having the non-lying original KantBot ruining countless surprise parties). Arguments of this kind are of great importance. But they can also make us blind to the overwhelming number of cases where either (i) normative theories give converging answers about what is morally right and wrong to do, or (ii) what we do doesn't seem to have any moral import.

One notable philosophical defense of (i) is the “triple theory” presented by Derek Parfit in *On what matters* (Parfit, 2011). According to this, the best and most defensible version of Kantian ethics, rule consequentialism, and contractualism of Scanlon can be described in a “single complex *higher-level* property wrong-making property, under which all other such properties can be subsumed, or gathered” (p. 414, original emphasis):

> *Triple theory*: An act is wrong just when such acts are disallowed by the principles that are optimific [productive of the best outcome], uniquely universally willable [Kant’s deontology], and not reasonably rejectable [Scanlon’s contractualism] (p. 413)

Parfit goes on to write: “It has been widely believed that there are such deep disagreements between Kantians, Contractualists, and Consequentialists. That, I have argued, is not true. These people are climbing the same mountain on different sides” (p. 419).

Now, it should be stressed that Parfit’s idea is controversial, and it would take a lot of work to explain more precisely how these traditions converge in a common ground, and why the interpretation Parfit offers of these traditions are the best and most defensible. It is, however, an intuitively plausible idea, and one that I was hoping to illustrate in the story with KantBot and Benthamizer. Following the upgraded KantBot, it seems that the most defendable Kantian duties are those that take into account the sort of beings that would find it possible to live according to those duties (being universally willed). To find them attractive, the beings would need to consider how the consequences of such duties affected their welfare (e.g., producing good outcomes). When the beings came together and discussed it, they also found that the best reason-giving moral judgements were those grounded in a mutual respect of others as reason-givers, which led them to formulate wrongness in terms of that others could reasonably reject to (contractualism). Relatedly, Parfit along with others believe that the most defendable forms of consequentialism – or, at least in terms of providing action-guidance – are those involving heuristics in the form of rules, standards, or principles that, if they were followed, would produce the most good. One reason is that, if we – like the original Benthamizer – tried to always produce the most good in every act we do (act utilitarianism), it could actually lead to worse outcomes than if we followed rules that tend to produce good outcomes (Parfit, 1984). As a simple example: we stop at a red light, not because the *act* of doing so in that specific moment creates the most overall good, but because we believe that the traffic *rule* promotes the most overall good.

While it might be at least plausible that normative theories converge at some abstract ‘higher-level’, there is a more practical sense of convergence that often goes unnoticed. The rough idea here is that moral dilemmas and disagreements are so entrenched in the conceptual DNA of what morality is that there is a sampling bias towards the times our moral intuitions disagree rather than agree. To this end, the very concept of morality inherently connotes conflict and choice – that one should do this rather than that – which creates a natural bias in our moral discourse toward situations of disagreement. However, these conflicts represent merely the 'tip of the iceberg' compared to the numerous instances where our moral reasons and intuitions actually align. It is only that we do not label those instances as 'moral'. This sampling bias overlooks the vast territory of implicit moral behaviors – those unconscious, intuitive, automatic, 'by default' – that form the foundation of our daily interactions. But because these behaviors already align with our implicit moral intuitions, we tend to categorize them as morally neutral or amoral, while elevating conflicts and disagreements to the realm of moral significance. This creates a somewhat misleading characterization of morality that overemphasizes discord while overlooking the broad foundation of convergence that underlies social behavior.

This practical point becomes more salient when we try to develop a morally competent machine, for when we do so, we first need to develop the appropriate 'baseline' of convergence upon which divergence arises. In other words, all the implicit moral behaviors that may be second nature and 'by default' for humans are not automatically so for machines, unless they are explicitly made that way. In the story, we saw how relatively basic algorithmic interpretations of normative theories yielded a significant divergence – for example, comparing the fast-acting KantBot with the slow-calculating Benthamizer. By contrast, the upgraded versions of Benthamizer and KantBot incorporated some form of decision-making heuristics of the other, leading to more convergence: KantBot considered the affected individuals, and Benthamizer employed more effective rule-based decision-making. For the same reason, I believe that the space of convergence will continue to grow the closer we get to the 'baseline' competence of a morally sensitive adult human being, just as it unlocks new forms of divergence from that base. In this view, it becomes easier to see normative theories, not as competitors, but as complementary decision-strategies for addressing a wider set of challenges. And in this picture, divergence is best understood as the few outliers that arise from the inconsistencies of this wider toolkit.

In Paper V, I discuss several forms of moral convergence in the context of moral machines along both theoretical and practical dimensions. One example is the analysis of general rules that incorporates sensitivity to others as part of their moral justification. In particular, I discuss different interpretations of the Golden Rule (GR), e.g., "treat others as you would like others to treat you". The GR is arguably the most famous example of a moral 'common ground', and it is often noted that formulations of it can be found in nearly all ethical and religious traditions (Blackburn, 2003). I argue that the best and most defensible versions of the GR are self-correcting, in the sense that they – like Kant's CI and Scanlon's contractualism

– involve a sensitivity to the ways in which one would like others to apply the GR, such as “Apply the Golden Rule to your general behavior in a way that you would want others to apply the Golden Rule to their general behavior” (V, pp. 33-35). This has the upshot that it can accommodate a great variety of moral elements emphasized by different traditions, such as respect, autonomy, sensitivity to outcomes, and rules. But it also becomes more demanding in another sense, as these elements must be mutually recognized among all agents, such as sharing the same moral expectations and ideas on what constitutes moral competence.

Another example is the divergence and convergence of outcome- and action-oriented theories, alongside the prospects and challenges for consequentialist-deontology hybrids as an approach to moral machines (V, pp. 56-59). In short, while there are philosophical, psychological, and technical considerations that make such hybrids promising, they inherit unresolved problems from both frameworks – for example, determining whether when to rely on established rules versus recalculating (a problem which the upgraded Benthamizer faced). Here, there is a convergence in that simple moral rules and complex moral calculations are two sides of the same coin, in the sense that, unless we blindly adhere to the former, their justification and successful application would still require the latter (I will discuss this form of convergence more extensively in lesson 7).

Lesson 6: From practice to theory

The sixth lesson contradicts the subtitle of this thesis, suggesting a shift from practice to theory rather than the other way around. As demonstrated throughout these chapters, strategies that begin with normative theory and proceed toward machine implementation face numerous challenges. But walking the opposite direction – starting with practice – not only offers a viable alternative but also circumvents many issues inherent to theory-first approaches. Here, I will outline how this approach might unfold.

First, we could start by examining a specific practice within a human domain – e.g., medicine, education, military, or transportation – where delegating morally significant decisions to computational systems is both ethically justifiable and desirable. This could, for instance, be a situation where we believe that such systems could reduce harm, comply to standards, or enhance moral goods, and possibly in ways humans cannot.

A practice-first focus constrains the scope of divergent moral machines in two important ways. (a) On a technical level, it provides numerous specific details inherent to the practice – about capacities, actions, environments, and resources – that inform the machine’s technical specifications. (b) On a moral level, it highlights normative elements tied to the practice itself. These need not necessarily have anything to do with normative theory, but could, for instance, encompass domain-specific ethical codes (e.g., medical or pedagogical ethics) or normative expectations of roles that have traditionally occupied that practice. A self-driving car might be modeled to emulate a smooth chauffeur, adhering to the subtle etiquette of local traffic norms. A social companion robot in elderly care might be modeled to capture norms associated with good care. The point is that practices already give a concrete structure that can inform both normative and technical features of the development, whereas theory-first approaches risk imposing generic solutions that overlook domain-specific nuances and challenges.

Next, insights from (a) and (b) can be used to clarify the cognitive capacities required for a machine to adhere to the technical and normative particularities of the practice. Here, we are naturally constrained by what is available and realizable by existing computational methods. In contrast, a theory-first approach might, in attempts to create a KantBot or Benthamizer, get stuck in infeasible Asimovian fantasies. Importantly, if what is feasible fails to meet the technical and normative specifications of the practice – e.g., fails to reduce the relevant harms or increase the goods that motivated its construction in the first place – then we should look for something else to do rather than building a machine.

If we take a practice-first approach, there is also a sense in which practical disagreements are constrained by their decision-making context in a way that theoretical disagreements are not. With regards to decision-making, the difference between practical and theoretical disagreements roughly correspond to the difference between, on the one hand, disagreeing about *what to do*, and on the other hand, disagreeing about the reasons or theories concerning *what to do* (e.g., based on what is morally good or bad).[24] We can think of a practical decision-making context as a range of possible actions with respect to a space of possible situations. An example of a minimal decision-making context is the Trolley Problem, which only involves one situation (a runaway trolley) with two possible actions (pull lever or not). Practically speaking, you either pull the lever or you do not – and this is all we can practically disagree over. Now, the practical decision-making contexts for an autonomous vehicle (AV) or a social assistive robot may potentially involve a vast number of possible actions and situations. But the practically relevant decisions – and the risks and costs they involve – are still *bounded* by the space of situations and range of possible actions of the decision-making context. Either the AV or social

[24] Arguably the most famous versions of this distinction can be found in *Principia Ethica,* where Moore distinguishes questions concerning “What kind of things ought to exist for their own sakes” from “What kind of actions ought we perform? (Moore, 1903, p. 2). Another version can be found in *Reasons and Persons*, where Parfit (1984) distinguishes between the moral ideal (concerning what is good) and the action-guiding components of normative theories (concerning what to do).

assistive robot do *this* or *that* – and this is all that we can practically disagree over. Theoretical and higher-order disagreements, by contrast, are not bounded in this fashion. When we disagree about the underlying reasons for doing this or that, about which ethical theory is correct, about what is good in itself, and of how to deal with disagreements and uncertainty about such questions, we engage in discussions that are potentially indefinite and unbounded.

The methodological reversal is explored in Paper II, III, and IV, and can also be found in broader design-frameworks such as *Value sensitive design* (Friedman et al., 2013) and methods for *Responsible AI* (Dignum, 2019). In II and III, I reframe virtue ethics, not as a 'top-down' or 'theory-first' approach to the development of moral machines, but rather as a generic blueprint of a cognitive architecture that can be adapted to accommodate practice-specific nuances in a bottom-up fashion. A central notion in virtue ethics is that of fulfilling one's function or purpose (*ergon*). In Papers II, I elaborate on how the practice-specific function a machine is meant to serve can help to determine what the relevant dispositions (virtues) the machine needs in order to fulfill that function well. Similarly, in Paper IV, after critiquing the 'theory-first' method, I suggest that we should "pay closer attention to the multifaceted ways normativity functions and serves human practices, and what capacities we should expect or even require participants in those practices to have" (p. 161). This, I suggest, does not necessarily mean that one need to get rid of normative ethics altogether, but rather, understand the futility of it unless it is grounded in the practices it aims to serve, and the cognition that is needed to successfully harness its benefits.

Lesson 7: Computational resources

The seventh lesson builds upon a point made in 3.3: that morality is hard. Fortunately, this hardness can serve as a point of convergence. The lesson is that, while normative theories diverge in the sense that they, as decision-making procedures, may capture different aspects of moral life, they converge in the fact that they utilize similar *resources* that are constrained in similar ways.

The most general and most commonly shared resource is that of time. Every decision problem that we face in everyday life requires some amount of time to make. We might spend some time deciding on what we are going to eat for lunch, whom to invite to a party, or where to go on vacation. Some decisions may be so difficult to make that, with limited time, we fail to make a decision at all. Therefore, in situations with critical time frames, it is important to – unlike the original Benthamizer – make up our minds before time runs out.

Another general resource that supports decision-making is that of knowledge. As with time, without any knowledge about what we like to eat, whom we enjoy hanging out with, or the appeal of different travel destinations, we cannot make informed decisions about what to order for lunch, whom to include on the guest list, or where to go on holiday. For the same reasons, when an emergency arises, it can be useful to have – like the original KantBot – quick access to the sort of knowledge that helps save those in need.

A third general resource is that of learning. It may be a bit odd to think of learning as a resource. Rather, it is perhaps better viewed as a combination of the first two resources, i.e., the ability to, over time, gain knowledge about something. With time, learning can help us figure out what we like to eat, which people we enjoy spending time with, and facts about European capitals we consider visiting. With the right kind of training, we can not only learn to act quickly, like the original KantBot, but adapt to changes and grow in light of new experiences like AristoMatic.

Interestingly, the decision-making of both humans and machines is bounded by these resources. Any mind or machine that attempts to do the right thing – regardless of how that is defined – is bounded by the time, knowledge, and learning it would take to successfully do so. In Paper V, I use computational complexity theory to explore in depth how causal engines, rule-followers, and moral learners (as described in section 2.3) are enabled and constrained by these and other resources. I also explore how the resources are interrelated. For instance, it might take a long time to make inferences based on a large knowledge-base, just as it might require a long time to learn certain complex things.

In computational complexity theory, a problem's complexity is defined by the resources an agent or algorithm requires to solve it. For computational systems, the two most critical resources are *time* and *space*. The latter typically refers to memory size (e.g., bits), while the former denotes the number of machine operations (or "state transitions"). The time and memory available to a computer at a given moment both enable and constrain the sort of problems it can solve, the knowledge it can store (e.g., in databases), and its ability to learn.

There is a general class of problems called P, encompassing all of the problems that can be solved in polynomial time. Formally, this means that the number of machine operations of an algorithm is upperbounded by (can *at most* use) a polynomial expression in its input, i.e., n^c, where n is the input size and c is a positive constant.[25] It is widely recognized among computer scientists that P captures an intuitive notion of problems requiring a 'realistic' amount of resources. For this reason, problems in P are called 'tractable', and those that require more resources – such as those whose runtime grows exponentially (c^n) – are called 'intractable'.

[25] In Paper V, I give an introduction to computational complexity that aims to be friendly to those who have never heard of it.

In Paper V, I demonstrate how a wide range of problems that normative theories face is intractable of this kind. In short, if moral decision-making includes planning, causal inference, dynamic and/or partially observable environments, strategic dynamics, logic, semantics, or learning, then moral decision-making presupposes intractable computations unless constraining factors are introduced.

One implication is that it presents an uncomfortable tradeoff between doing what is morally right (according to the prescriptions of the normative theories) and doing what is feasible. For emergencies, we would not want a machine that gets lost in calculations like Benthamizer. But such exhaustive calculation may be suitable for contexts with ample time, such as coming up with fair and efficient distribution plans. These computational constraints, I argue, can help us identify the situations and contexts in which certain forms of action-guidance can be expected to be more viable than alternatives.

A more intriguing implication is that human morality faces analogous constraints, which can in turn be used to illuminate human moral cognition. In Paper V, this idea is captured in the Moral Tractability Thesis (MTT). It states that moral behavior, problem-solving, and cognition are constrained by computational tractability, given some reasonable model of human moral cognition. Now, we do not know what kind of computer the human brain is, or whether it can be meaningfully characterized by any model of computation. But what we do know is that no human mind seems able to solve intractable problems in ways that violate widely believed conjectures in computer science, such as P ≠ NP (where NP is the class of problems solvable in polynomial time by a nondeterministic Turing Machine).[26] In this way, MTT can provide a constraining factor for existing paradigms studying human morality, from the computational modeling of human moral cognition to experimental studies in moral psychology. More specifically, MTT can be used to identify the principles and algorithms that underpin cognitive processes in moral decision-making, reveal relevant trade-offs between feasibility and performance, and further investigate the role of resources in specific contexts.

In Paper V, I also use the resource analysis to spell out more specific lessons about the computational nature of moral decision-making. Here, I will expand on three: (i) the complex simplicity of rule-following, (ii) challenges of learning, and (iii) strategic dynamics. I believe each lesson tell us something interesting about moral decision-making for machines as well as humans.

(i) ***Rule-following.*** There is a widespread conception – among machine ethicists, moral psychologists, and moral philosophers – that rule-following is, in some significant sense, a more practically viable method for moral decision-making than

[26] In cognitive science, this point is reflected in the P-Cognition thesis, which asserts that human cognitive functions are constrained by polynomial time (Van Rooij, 2008).

its alternatives (see Paper V, pp. 29-30). Intuitively, rules of the form “do not X” are easy to follow and understand. Such considerations allow us to distinguish normative theories as ideals and decision-procedures (Parfit, 1984). With respect to the latter, rule consequentialism (part of the upgraded Benthamizer) is often believed to have a clear edge over act utilitarianism (the original Benthamizer). In moral psychology, similar considerations have been key in the development of the dual process theory of moral cognition (Greene et al., 2001), which posits that moral judgments rely on both ‘slow’ conscious-controlled processes (corresponding to typically utilitarian judgements) and ‘fast’ automatic-emotional processes (corresponding to typically deontological judgments). The view is also prevalent in machine ethics, where one may think that deontological rules of the form “If input X → do Y” elegantly reflect the sort of conditional statements ubiquitous in machine code.

In section 5 of Paper V (pp. 29-61), I argue that this conception of rule-following is misguided, as the simplicity of rules can only be secured in either complex or controversial ways. In brief, here are five aspects of this complex simplicity:

(a) While rule-following in human contexts may appear simple (e.g., laws), they rely on a complex relationship between the mechanisms that incentivizes their adherence (e.g., punishment and police), ensures their just interpretation (e.g., courts), and the moral views of the subjects the rules apply to (e.g., citizens).

(b) Alternatively, while rule-following justified on the basis of divine command or legal positivism may avoid some of these difficulties, the knowledge-requirements that are needed for them to work in practice can be unfathomably vast (e.g., knowing whether a rule applies in a specific situation), while the knowledge itself, and the justification for it, will be highly contentious.

(c) Another alternative is for rule-following to include some form of methods for justification as part of the rule itself, as exemplified by Kant’s Categorical Imperative, Scanlon’s contractualism, or certain variants of the golden rule (lesson 5). However, the application of such self-justifying rules is obfuscated by extreme variances with regard to behavioral expectations, preferences, and cognitive capacities of other agents in heterogeneous populations (and for the same reason, such rules would thrive in perfectly homogeneous societies, e.g., in an ideal Kingdom of Ends or in a Hobbesian war of all against all).

(d) I also demonstrate that any computational system that employs formal logic, e.g., for rule-following, moral reasoning, or communication are subject to expressibility, intractability, and decidability results that permeate the syntactics and semantics of logic.

(e) Finally, I also explore how rule-following can enhance the run-time efficiency of consequentialist decision-making, with the tradeoff that such efficiency is either based on optimistic conservatism ("it has worked before so it should work again", or "the rule is there for a reason, even if I don't know which"), the results of some other complex process (e.g., reasoning or learning), or the collection of vast moral knowledge. I use these insights to ultimately conclude that, while the power of rules lies in their general applicability, general justification, and computational simplicity, it is a power that can only be secured in complex or problematic ways.

(ii) *Learning.* Like rule-following, learning may at first glance seem to make everything easier, at least post-training. But to effectively and rigorously learn something, you need lots of training time along with a good quantity of high-quality training examples (i.e., the training examples are representative of what you actually want to learn). Like AristoMatic, you may need lots of examples of recklessness and cowardice to learn what manifests proper courage. Some learning processes – e.g., that of exploring a wilderness full of dangerous creatures – can themselves introduce massive risks, as one mistake can put an end to the learning itself.[27] You, like all modern learning systems, may need some form of inductive bias. Inductive biases are assumptions – e.g., about the problem you want to tackle, the problem space, the distribution of data, the quality of data, etc. – that we can exploit to enable and foster learnability.

In section 6 of Paper V, I explore in detail the sort of problems that learning leads to (pp. 59-69). One is the possibility of bad distributions of data. This relates to an impossibility result known as the "No Free Lunch Theorem" (NFL), which establishes that there will always be unfortunate distributions for which the *sample complexity* – the number of training examples required to learn a target function – is arbitrarily large (pp. 63-64). It implies that there is no learning algorithm that can perform well on every learning task having trained upon a dataset of a fixed size.[28] As a solution, however, we might have model-relative justifications in order to explain why some inductive inferences seem to work better than others (inductive biases) or constrain the space of hypotheses in favor of the simpler (Occam's razor). But both alternatives lead to other problems. There is no guarantee that our inductive biases capture the world out there, or capture it in the way we hope to, just as there is no guarantee that the simplest hypothesis gives the best explanations. In practice,

[27] Many computational learning techniques, such as reinforcement learning and stochastic methods at large, presuppose trial-and-error. Although this might not be an issue in virtual environments, it presents challenges for real-world environments where exploration may not be viable, such as that when some actions have catastrophic consequences.

[28] Interestingly, this can be related to Hume's problem of induction (Sterkenburg & Grünwald, 2021). Hume famously advanced skepticism against the very justification of induction, arguing that deductive reasoning alone cannot secure the validity of inductive inference; and neither can induction, due to circularity, provide non-deductive grounds for itself (Hume, 1739).

however, what is surprising is not the general trend of results showing that learning is challenging. It is rather that we lack rigorous explanations for *why* some learning systems seem to generalize well in practice. This is known as the "paradox of deep learning", which centers around understanding the empirical success of deep learning despite the absence of theoretical explanations (p. 66).

One problem of inductive biases that deserves special attention is the case of language models, most famously associated with how they power popular services like ChatGPT (OpenAI), Gemini (Google), LlaMa (Meta), and Claude (Anthropic). In a nutshell, a language model is a probabilistic model of natural language that, having trained on a text corpus, can generate probabilities of a string of words based on the text corpus. In recent years, language models have become so big – with hundreds of billions of parameters, hundreds of billions of training data points, and hundreds of years of training[29] – that we refer to them as Large Language Models (LLMs). Their most notable aspect, however, is not their size but their capabilities for general-purpose language generation and understanding. Studies have shown LLMs solving Theory of Mind tasks (Kosinski, 2024) and passing the bar exam (Katz et al., 2024). More interestingly for this thesis: LLMs have been able to replicate human moral judgements with near-perfect accuracy (Dillion et al., 2023), tailor their moral judgments based on a user's political identity (Simmons, 2022), and even outperform expert ethicists in giving trustworthy and thoughtful moral advice (Dillion et al., 2024).

Despite these impressive results, the training and development of LLMs raise a host of convoluted questions about inductive biases. One issue concerns the training that goes into the LLM. For instance, how should we consider the moral competence of an LLM given that its training data (the internet) – is deeply imbued by (potentially) problematic biases (Liang et al., 2022)? This assumes that the internet is an accurate representation of human values and ethical thought – an inductive bias one might reasonably question. It may also present an uncomfortable compromise between helpfulness and harmfulness. While massive amounts of unlabeled training data – hundreds of billions of byte-pair-encoded tokens – may be required to support the helpful text-processing capacities of an LLM, it inevitably includes undesirable behaviors that can at best be inhibited, yet never be avoided altogether (Wolf et al., 2023). Another issue relates to the "paradox of deep learning": how do we understand requirements of transparency, explainability, robustness, safety, and fairness for sufficiently advanced 'black box' systems (Paper V, pp. 68-69)?

[29] For instance, even the now 'old' GPT-3 (released 2020) has 175B parameters and was trained on more than 410B byte-pair-encoded tokens over an estimated 355 years single GPU time (Brown et al., 2020). While only certain technical data has been released about its successor GPT-4 (at least by the time of writing this), it is rumored to have around 1.7 trillion parameters (Hudson et al., 2023).

It is possible to alleviate some of these issues by fine-tuning the black box. One commonly used technique for this purpose discussed in lesson 4 (incomparability) is RLHF. It is hypothesized that RLHF is the secret juice that makes ChatGPT superior – being far more user-friendly and exhibiting less reasoning biases – to the LLM it is based on, such as GPT-4 (Hagendorff et al., 2023). RLHF works by training a reward model based on human feedback – e.g., a preference ranking of the outputs generated by the system – which is then used to fine-tune the model (Ziegler et al., 2019). This, however, begs the question: who are these humans giving feedback, and what and whose values do their preferences represent? And as we saw in lesson 4, aligning AI models with the preferences of everyone will inevitably violate the preferences of certain individuals.

As I explore in Paper V, the success of learning systems (e.g., in terms of training efficiency and predictive accuracy) seems inversely proportional to the inductive assumptions they exploit. That is, for moral learning to work, we need to have a relatively clear idea of the performance measure – e.g., in terms of some predefined score, goal, or objective function – for the problem space we want the learning system to tackle. Similar to divine command and legal positivism, this presupposes that we have an answer to the questions we seek. Alternatively, we may – as in the case of LLMs – simply hope that it exists in the vast statistical ocean of the training data, or that the currents in this ocean can be fine-tuned in the appropriate ways (with techniques such as RLHF). Thus, the lesson of moral learning does not reside in the computational complexity of learning as such, but rather in justifying the assumptions we need to exploit in order for learning to work, or accept that it somehow works, but we don't know why.

Here, it is worth emphasizing how "time flies in the age of machines" (section 1.4), and that the theoretical results discussed in Paper V should be taken with a pinch of salt in light of the recent explosion of practical machine learning developments and applications. This rapid pace makes it incredibly difficult to pin down firm conclusions about the capabilities and limitations of machine learning systems – or to make reliable predictions about their future development. As highlighted by the paradox of deep learning, AI development is progressing faster than our ability to fully understand it. Traditionally, as I discuss in Paper I, computer science was a theoretical and analytical field. But in the era of machine learning, it has become increasingly empirical. This shift means that for most advanced AI models, there are no clear-cut mathematical or logical guarantees of performance or robustness – just empirical test results based on benchmarks. This means that, in many ways, we're building the plane as we fly it, all while trying to figure out how it stays in the air.

(iii) ***Strategic dynamics.*** Many interactions we face in everyday life have some kind of strategic logic to them. When someone asks for your help, you might think, "what's in it for me?". Some interactions might involve salient benefits for all parties, such as in a fair trade of goods where everyone ends up with what they want. Sometimes the benefits are less direct or merely potential. We might help an acquaintance move with the promise that they will, at some point in the future, help us move. Sometimes we are lied to or stolen from, and those who lied to or stole from us probably did so because it benefited them, perhaps knowing that the chances for retaliation are slim.

That was a rough and informal introduction to game theory, the mathematical study of strategic interactions. If you ask the biologist from section 3.1, they might give you an elaborate story about how evolutionary game theory can show how cooperation – and perhaps morality itself – arises and persists as a logical consequence of strategic dynamics among self-interested individuals (Axelrod & Hamilton, 1981; Nowak, 2006). This is far from a new idea. Long before the formal foundations of game theory were developed by John von Neumann, Oskar Morgenstern, and John Nash (Nash, 1950; von Neumann & Morgenstern, 1944), philosophers such as Hobbes (1651), Hume (1739), and Rousseau (1755) used informal game-theoretic arguments in their influential analyses of morality.[30]

In 5.1.3 of Paper V (pp. 35-45), I get into more detail of how complicated things can get if our moral decision-making needs to account for strategic dynamics.[31] In particular, I explore the benefits, drawbacks, and challenges of pure strategies (e.g., either do the altruistic or the self-interested action) and mixed strategies (seeking to maximize self-interest), in situations of complete and incomplete information (with strategies maximizing *expected* self-interest), and the challenge of computing morally attractive equilibria – where no player can gain anything by changing their strategy – in various settings (they turn out to be hard!).

One surprising finding is that cooperation can sometimes be the result of computational constraints, such as a small memory, restricted information about others, or that of being satisfied with 'close enough' (Paper V, pp. 36-37). Intuitively, short-sightedness could help us forgive and forget the misdemeanors of others. A similar observation can be found in the case of recursive reasoning about higher-order beliefs (i.e., beliefs about others' beliefs about one's beliefs …), which, while being cumbersome to compute, provides no significant strategic advantages beyond recursion levels higher than 2 (pp. 38-39).

The broader lesson is that any form of moral decision-making that seeks to account for the strategic interactions that occur when living with others – who may have

[30] See also Gauthier (1986) for a more contemporary work along this direction.

[31] In fact, the first publication in this project – later on excluded from the thesis – explored how *persistent homology*, a computational technique based on abstract algebra, can be used to study spatial game-theoretic dynamics (Stenseke, 2021).

different strategies and preferences – must take game-theoretic considerations into account. Earlier, I claimed that the real-world moral environments we find ourselves in are of this kind, presenting a complex, heterogeneous mix of agents and games.

The generality of these lessons should be emphasized. Any machine or human that employs some form of rule-following, learning, or strategic reasoning in their moral decision-making faces, in some way or another, the aforementioned challenges. A similar lesson can also be spelled out about causal inference, as even if one is not a consequentialist, one may still think and act in a world of causes and effects. But I would rather advice you to read section 3 in Paper V to learn this lesson yourself.

In sum, focusing on the computational resources that goes into moral decision-making opens up interesting interdisciplinary venues between machine builders, moral philosophers, and moral psychologists. I believe it can inspire new directions, not only in the engineering of moral machines, but in understanding the rich and multifaceted nature of morality.

Lesson 8: Cognitive capacities

While moral theories diverge in that they, as decision-making procedures, may utilize different cognitive capacities, they converge in the fact that these capacities are often shared by those who participate in such procedures. Moral life may be heterogeneous and complex, but human moral cognition faces it as a unified whole. Drawing from Paper IV, the upshot is that, if we carry out our moral machine-building project correctly, there is a possibility for machines to have these capacities too. Just as time, knowledge, and learning are fundamental resources for moral decision-making, so too are the cognitive capacities that make use of them.

Lesson 9: Unifying frameworks

The final lesson is that convergence can be achieved by exploring general frameworks that can help to reconcile the divergences. In this work, I have mainly explored four such unifying frameworks: virtue ethics, computational complexity, game theory, and philosophy of science. What they share is their ability to (a) integrate insights from different disciplines and (b) offer common ground for different (and potentially conflicting) conceptions of morality, such as diverse moral traditions and normative theories. It should be stressed that these are not the only frameworks that can serve as unifiers in navigating the landscape of morality; rather, they are the ones I have had the pleasure of exploring in greater depth over the past four years.

In Papers II and III, I argue that virtue ethics offers an interdisciplinary-apt framework that illuminates what it could mean for a machine to have a moral character. It should be noted, of course, that specific versions of virtue ethics may conflict with other normative theories, in particularly regarding central emphasis on virtue (Hursthouse & Pettigrove, 2023). However, independent of its more substantial normative prescriptions, virtue ethics offers valuable insights into what it means to be a moral being and to flourish as such.

In the papers, I describe three ways in which virtue ethics can serve as a unifying framework for moral machines: (a) the capacity to leverage modern machine learning techniques via the notion of practical wisdom, (b) its provision of a more holistic picture of moral cognition that resonates with insights from psychology and neuroscience, which together can yield (c) flexible architectures that accommodate the importance of features of both rules and consequences (moral convergence).

In Paper V, I describe how computational resources both enable and constrain moral decision-making in minds and machines. In Papers II and V, game theory is presented as a framework for analyzing, describing, and evaluating aspects of moral decision-making. Finally, in Paper I, I describe how insights from philosophy of science can be used to integrate multiple disciplines. Specifically, I propose 'metacognitive scaffolds' as an epistemic resource to articulate and analyze how a given discipline, with its discipline-specific beliefs, methods, and values, generates and applies knowledge. I then describe how such scaffolds can elucidate disciplinary perspectives relevant to machine ethics.

4.3. Summary

In sum, to reach the robot lab, one must pay close attention to the divergences and convergence within morality. On the one hand, challenges arise from particularities (lesson 1), interpretation (2), technical details (3), and incomparability (4), which together open a vast space of divergence. Within this space, it is easy to lose one's way or lost or abandon the project altogether. Yet, within this same space, points of convergence also exist – points of agreement and synergy (5), practice-specific anchors (6), shared resources, constraints, and cognition (7 and 8), knitted together by unifying frameworks (9) – all of which can make the space easier to navigate. Even if these points are insufficient to chart a direct path to the robot lab, I believe they have at least offered interesting, and perhaps novel, insights into morality.

Here concludes the focus on the development of *explicit ethical agents*. In the next chapter, we will turn to the topic of *full ethical agents*, which may also possess the metaphysical properties essential to human moral agency, such as consciousness and free will.

Chapter 5 – Natural and artificial moral agency

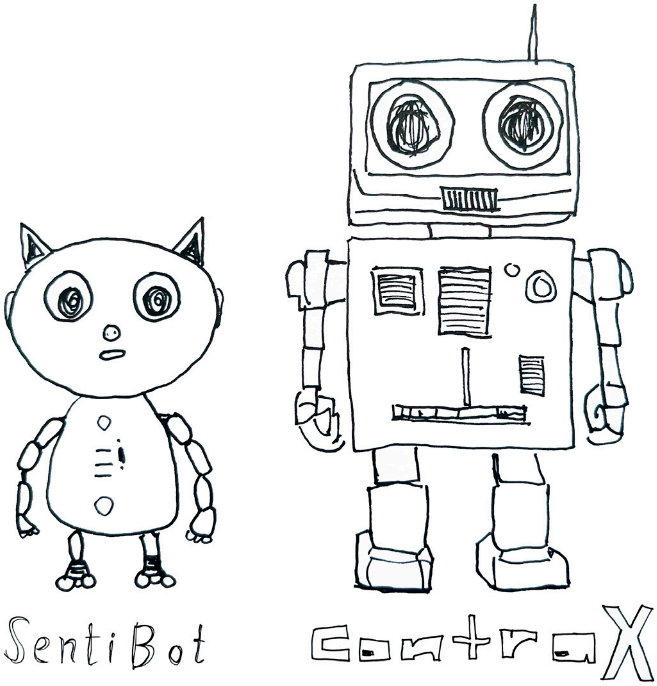


Figure 4: Two *really* nice robots.

5.1. A tale of two moral agents

Another day, another mysterious robot lab on the top of another hill sent two boxes to another town with a note: “We gift you these two moral robots that will make your life flourish”. Intrigued, the townspeople gathered to unpack SentiBot and ContraX, unaware of the philosophical debate they were about to ignite (Figure 4).

SentiBot, with its lifelike expressions and emotive responses, immediately endeared itself to many. Based on David Hume's naturalistic philosophy, it possessed evolved dispositions and sentiments that formed the basis of its moral judgments.[32] When a child fell and scraped their knee, SentiBot rushed to comfort them, its eyes reflecting genuine concern. "I feel your pain," it said, gently patting the child's back. This display of empathy resonated deeply with those who believed that morality stemmed from our emotional capacities as living creatures.

ContraX, in contrast, stood tall and impassive. Though not sentient like SentiBot, it boasted sophisticated moral reasoning capabilities rooted in the contractualism of Scanlon and deontology of Kant. When faced with a complex dispute between neighbors, ContraX methodically analyzed the situation, referring to pre-existing agreed-upon rules, or in cases where there were none, principles that everyone would deem reasonable if they thought about them carefully. Its approach appealed to those who viewed morality as a product of reason and social contract.

The town's divided opinion on the robots' moral agency was put to the test when both made errors in judgment. During a town fair, SentiBot accidentally knocked over a display of delicate artwork while trying to help a lost child. Immediately, its face blushed with distress. "I'm so sorry," it said with a trembling voice. "I feel absolutely mortified about this. How can I make it right?" Its display of guilt and remorse moved many observers, who saw it as a genuine sign of moral awareness.

ContraX, tasked with optimizing the town's budget, made a miscalculation that led to a shortfall in the education fund. When confronted, it responded with cool precision: "Mistakes have been made. I accept complete responsibility. Let us analyze the error and implement corrective measures to prevent future occurrences." Its lack of emotional response disturbed some, while others appreciated its focus on rectification and future prevention.

As the robots integrated further into the community, people began to discuss which one truly embodied moral agency. Some argued that SentiBot's emotional ability to resonate with the raw and messy aspects of human nature made it a genuine moral agent. Others contended that ContraX's rational, unbiased decision-making was superior, being more reliable and truly ethical.

Sarah, a local psychologist, argued passionately for SentiBot. "Morality is fundamentally about feeling" she insisted. "SentiBot's ability to emotionally connect with others is what makes it a true moral agent. It's not just following rules. It's genuinely caring."

[32] SentiBot should not be read as a literal embodiment of David Hume's moral philosophy, which, besides the role of innate sentiments (e.g., feelings of sympathy) also involves the 'artificial virtues' needed for successful impersonal cooperation (e.g., justice).

Mark, a philosophy professor, countered, “But isn’t morality about rising above our base instincts? How else are we humans different from animals in the wild? ContraX’s capacity for impartial reasoning and adherence to universal principles is the hallmark of true moral agency. It’s not swayed by emotions that might cloud judgment.”

Amidst the disagreement, an elderly resident named Alice spoke up. “Perhaps,” she suggested, “true morality lies not in choosing between sentiment and reason, but in finding a balance between the two. SentiBot reminds us of our shared human heart, while ContraX ensures we make use of our rationality.”

Another resident, a young snarky teenager named Lukas, claimed that neither SentiBot nor ContraX were moral agents on the basis of the simple argument that, “machines cannot simply be so”. He presented an eerie possibility, “SentiBot only acts as if it is empathizing with us, but do we actually know this? What proof do we have beyond its observable behavior? For all we know, it could just be pretending to feel”. Sarah, the psychologist sympathizing with SentiBot found this deeply disturbing, “and what proof do we have beyond observable behavior that *you* are not pretending, Lukas?”

As the sun set on the hill, the town had not reached a consensus on which robot was the true moral agent. Instead, they had embarked on a journey of moral inquiry, grappling with fundamental questions about the nature of ethics, and the role of emotions and reasoning in decision-making. The mysterious lab’s gift had indeed made their lives flourish, not by providing easy answers, but by inspiring deep reflection on what morality is, and what it means to be a complex creature in a world of other complex creatures.

******

The story is meant to juxtapose two traditions of how we might think of what makes us “full” moral agents, drawing on a common (and perhaps problematic) dichotomy between thinking and feeling. There are, of course, other traditions and hybrid approaches beyond this binary. Interestingly, under the least restrictive views, some current AI systems might already qualify as ‘full’ moral agents. As we saw earlier, LLMs have been shown to replicate human moral judgements (Dillion et al., 2023), manifest moral flexibility (Simmons, 2022), and may soon threaten to put expert ethicists out of their job as moral advisers (Dillion et al., 2024).

Yet, my guess is that most of us think that this is not enough for genuine moral agency. No matter how perfectly a robot like ContraX can learn to navigate the intricacy of human preferences through RLHF, or by fine-tuning its own behavior via human-defined moral principles through Reinforcement Learning from AI

Feedback (RLAIF),[33] something seems missing. Like the skeptic in the section 2.2, one might argue that consciousness is necessary – a quality ContraX lacks. If so, the question of artificial moral agency becomes a question about the possibility of artificial consciousness, an inquiry which in turn leads to tricky questions about the nature of consciousness.

In this chapter, I will dig these debates and my contributions to them. I will begin by reflecting on what makes 'full' or 'genuine' moral agents different from the machines discussed in previous chapter. I will also distinguish moral agency from another central concept – that of moral status (5.2). Next, I will examine moral agency in terms of *capacities* that an agent must have in order to qualify as moral (5.3). I will focus on three such capacities – rationality, autonomy, and consciousness – asking whether AI systems *can* have them, and whether they *should* have them. Finally, I problematize the strategy of approaching moral agency in terms of capacities by considering the ways in which the concept is shaped by how agents relate to each-other in virtue of relationships and practices (5.4).

5.2. Moral agency and moral status

If we, as a starting point, define a moral agent as "someone capable of acting based on some notion of what is morally right and wrong", this would include the 'explicit moral agents' discussed in the earlier chapters. However, this definition overlooks a subtler, more elusive concept. This can be clarified by considering the following scenario.

Imagine programming the accident-management system of an autonomous car to minimize harm by colliding with the fewest people possible. In a situation where the car must choose between either (a) continuing its course and hitting two pedestrians, or (b) swerving into another lane and hitting one pedestrian, it selects option (b). Suppose we programmed it like this for moral reasons, believing that minimizing harm is morally right. If so, we might conclude that the car is a moral agent according to the initial definition: it is acting based on some notion of what is morally right.

[33] RLAIF, or Constitutional AI, is another LLM fine-tuning technique, which centers on providing human supervision in terms of a list of principles or rules (i.e., a constitution). In short, the system generates responses based on harmful prompts and receives critique of those responses in light of the constitution. The critical feedback is then used as feedback to revise the system's own responses (Bai et al., 2022).

But isn't it then acting in virtue of the programmer's notion of what is right? After all, at no point does the car make the decision *itself*, it is simply following a code based on a decision already made for it.

Another consideration that can cast doubt on the moral agency of the car is that of responsibility, which, at least for human moral practices, is a cornerstone of what it means to be a moral agent.[34] If you are a responsible moral agent, it would be appropriate for me – another moral agent – to blame you for your wrongdoing. But can we say that the car is responsible, and blame it for hitting the one pedestrian in the example above? This seems misguided. Rather, perhaps one would hold that we (the programmers) are responsible for programming the car in this or that way, or that the user – being transported by the car – is responsible.[35]

From this, we can derive two preliminary desiderata that distinguishes 'genuine' moral agents from the machines discussed in the previous chapters. The first is related to some choice or control that the agent has over their decisions, which cannot simply be the result of someone else forcing or making it do so (e.g., programming). The second is the idea that it should make sense to hold the agent *responsible* for making its choices.

Certain choices that responsible agents make can in turn be reacted to. Typical reactions are that of blame and praise, the former expressing a normative significance towards the agent's behavior (or the agent themselves) of a negative sort ("It was wrong to do that!"), the latter of a positive sort ("Well done!"). A moral community may be filled with a variety of these other-directed responses, accompanied by emotions such as anger, disgust, and resentment. Some agents may also react to their *own* behavior, with self-directed emotions such remorse, guilt, and shame. Surely, people might blame each-other for all sorts of reasons. But often, there is a hope that our reaction has some form of impact on the agent, signaling that they should "think more carefully before doing that again!".

This is one possible explanation for why it seems absurd to blame the car: it doesn't have the capacity to recognize the emotional reactions we direct towards it in the appropriate ways. And if we find ourselves in camp SentiBot, this suggests a third desideratum. We may hold that the emotional nature of these reactions, that they are

[34] During my time as a doctoral student, I have learned heaps about moral responsibility from colleagues in the Lund Gothenburg Responsibility Project, which includes the excellent dissertations by Velichkov (2023), Werkmäster (2023), Emilsson (2024), and Mirzaeighazi (2024).

[35] In fact, buried in these sentences is a vast discussion in AI ethics about the responsibility of autonomous machines, referred to as the 'responsibility gap'. See Sparrow (2007), Champagne and Tonkens (2015), and Mirzaeighazi and Stenseke (2024) for three different views.

felt or *experienced*, along with the moral awareness of others and oneself they entail, plays an essential part of what it means to be a moral agent.

For ContraX advocates, however, this desideratum may be of less importance. What is essential, instead, is that we are able to rationally reflect upon our interaction with others and decide on how to act in virtue of this. Surely, emotions can play an important part in this, but emotions are messy, and on occasion, a rather bad guide. Like a programmed car, emotions can be seen as forcing us to do or feel things beyond our control. The whole point of morality, we might say, is about the part that we actually *can* decide, not in virtue of evolved dispositions that were given to us (like the programmer giving code to the car), but in virtue of the uncoerced autonomous thinking we have as free rational creatures. Like Mark, we could note that other animals also have emotions, yet, their inability to act rationally makes them incapable of acting morally. This line of reasoning yields another desideratum for moral agency: a capacity for rational thought and action, which is distinct from the capacity to merely feel.

To sum up the discussion so far, a moral agent is someone capable of making their (i) *own* decisions in virtue of some notion of what is morally right and wrong, and this 'own'-making process is grounded in either (iia) their emotional capacities (e.g., phenomenal experience or moral awareness), their (iib) rational capacities (e.g., for autonomous thought, and to act based on reasons), or both. Furthermore, (iii) it should not be odd to hold a moral agent responsible for their decisions, plausibly in virtue of either (iia) or (iib). In the next section, these will be analyzed in terms of *capacities* required for full moral agency, where (i) will be described in terms of *autonomy*, (iia) in terms of *consciousness*, and (iib) in terms of *rationality*.

But first, it is important to distinguish moral agency from moral status.[36] If a being has moral status, it means that there are moral considerations and obligations that applies to how we – moral agents – treat that being for its *own* sake. It is widely believed that all sentient beings have some form of moral status (Jaworska & Tannenbaum, 2023). When we recognize a being's sentience – i.e., the ability to experience sensations such as pleasure and pain – we may feel obligated to take its welfare into account in our moral deliberations, such as avoiding causing it unnecessary harm. This fundamental respect for sentient organisms forms a cornerstone of many ethical frameworks, acknowledging that the ability to feel is itself a morally relevant characteristic that demands our attention and care.

[36] Terminological note: in the literature, 'moral status' is often – but not always – used synonymously with terms such as 'moral standing' and 'moral significance', and a being with moral status is often called a 'moral patient'. In this section, I will abide to this convention.

Moral agency and moral status are intertwined in complex ways. I will describe two. First, we note that moral agents are those that have moral considerations for beings with moral status. If a being is perceived as having moral status, it is so in light of a relationship to (some) moral agent(s). This means that whatever reasons moral status yields, they must, in some way, manifest in the minds of the moral agents who act upon them – e.g., by treating those with moral status in certain ways. For instance, it is generally believed that adult humans are both moral agents and moral patients, while human infants and non-human animals are moral patients but not moral agents. In practice, this means that the moral considerations of human infants and animals are so to speak 'in the hands of' the adult humans who act with moral consideration towards them. This is a point expanded upon in Paper VII, and something we will return to in section 5.4.

The second is that properties often viewed as necessary for moral status are closely associated with properties viewed as necessary for moral agency. According to some views, there are (a) capacities that makes one a moral agent (e.g., autonomy) that provide grounds for moral status (or specific kinds of). According to other views, there are (b) capacities associated with moral status (e.g., sentience) that provide grounds for (certain forms) of moral agency.

Let's start with (a). Perhaps the most famous example is Kant, who argues that only autonomous beings – i.e., those acting in accordance with their own self-imposed rules – should be treated as ends in themselves, which is also what grounds a person's dignity (Kant, 1785). More recently, Shelly Kagan (2019) has argued that "agency of any sort suffices for moral standing of some kind" (p. 30, original emphasis). To support his claim, Kagan imagines a planet of advanced robots that, although sharing various aspects of human civilizations – they reproduce, have culture, belong to communities, have sophisticated plans, etc. – they lack sentience. Kagan argues that these robots still have a moral standing: for example, they may (following contractualism) rationally act in accordance with agreed-upon principles (e.g., forbidding one to kill robots), and they may – following Kant – pursue their autonomously set goals.

Now an example of (b). Quite forcefully, Himma (2009) argues that standard accounts of moral agency all implicitly presuppose consciousness, which many takes as sufficient and/or necessary for moral status (Jaworska & Tannenbaum, 2023). One argument Himma gives is that it is a conceptual truth that the actions of a moral agent are the result of intentional states, which are *mental* states. Another is that, without a first-person conscious perspective, a being would lack agency, as there would not be a perspective from which the agent acts. A third argument is that – like the car example, or perhaps ContraX – it would be odd to react with praise or indignation to something without conscious states (the responsibility desideratum based on a capacity for having subjective states, such as emotions).

In Paper VI, I analyze both the moral status and moral agency of artificial beings in light of Kazuo Ishiguro's novel *Klara and the Sun* (Ishiguro, 2021). The analysis illustrates how moral status and agency are intertwined in the two ways just described. I juxtapose two approaches to moral agency and status which I call the "view from within" and the "view from outside". The first centers on Klara's first-person perspective: what it means to have moral status and to make moral decisions based on one's own subjective point of view – the "what it is like" the be conscious, autonomous, and rational. The second centers on how other characters assess the moral qualities of Klara based on her outward behavior. I argue that the novel exposes the shortcomings of the "view from within" in relation to the social reality imposed by the "view from outside". That is, regardless of what moral qualities we can attribute to artificial beings like Klara "from within", these will ultimately be determined by the views of others "from outside". The interesting – and somewhat convoluted – upshot is that other's views are not restricted to externally observable behavior, but can also involve metaphysical ideas the others have about the nature of consciousness and personhood. Building upon this inquiry, in Paper VII, I offer an account that seeks to reconcile the tension between the "inner" and "outside" view (the details of which will be discussed in 5.4).

My reading of Ishigoro's novel has led me to believe that moral agency and moral status are closely linked, and that any account of either must include an account of the other. In the rest of this chapter, however, I will bracket this thought and focus primarily on moral agency.

5.3. Rationality, autonomy, consciousness

In the previous section, we identified three preliminary criteria for what constitutes a full moral agent: someone capable of making their own (autonomous) moral decisions, based on their capacity for rational thought (rationality) and subjective experience (consciousness).[37] I will henceforth refer to these three capacities collectively as RAC. This raises the question: can machines possess RAC in a manner that qualifies them as full moral agents?

As we reflect on the future possibilities of artificial intelligence, it is important to approach this question with humility. We must recognize both our inability to predict the future and the long-standing, scientifically and philosophically contested nature of mental capacities like RAC.

[37] See Paper VII for a more detailed exposition of the 'capacities view', with references to thinkers that have, in various ways, advanced claims about capacities that are necessary for moral agency.

I must also confess that my own views on this topic have been consistently inconsistent, and that traces of this flux can be found across the seven papers that comprise this thesis. In what follows, I will draw from both historical and contemporary debates to outline three key challenges to the prospect of artificial RAC – challenges that may help explain my own evolving stances.

Challenge 1: The ocean between mind and matter

Few topics in the history of philosophy have generated as much enduring debate as the nature of consciousness. A classic formulation is the mind-body problem: how the mental phenomena of thoughts, feelings, and experiences relate to physical states and processes of the brain and body – whether they are identical, separate, or connected in some other way.

René Descartes (1637) famously encapsulates one contribution to this problem with his dictum *cogito ergo sum* ("I think, therefore I am"). Descartes thought that the very act of doubting one's existence confirms the reality of the doubter – an undoubtable certainty rooted in the subjective experience of thinking. If you meditate on this, you might, like Descartes, arrive at a fundamental division between mind (res cogitans, or 'thinking thing') and matter (res extensa, or 'extended thing'). This could turn you into a dualist, who maintains that mental phenomena possess a non-physical quality that cannot be accounted for by physical processes alone.[38]

In contrast, physicalist views hold that consciousness is fundamentally physical in nature. The physicalist could hold that mental states are simply *identical* to brain states, just like water is identical to H_2O (type identity physicalism). A more radical specimen of physicalism holds that our intuitions about mental states (such as beliefs or desires) are fundamentally misguided: they do not actually exist – at least not in the way we commonly think about them – and will eventually be replaced with more accurate neuroscientific descriptions (eliminativism). A more moderate specimen is functionalism, which defines mental states in terms of their functional (causal) roles, as opposed to the physical substrate they are implemented in.[39]

The tension between these views persists not only in contemporary debates in philosophy of mind, but across neuroscience and AI. For instance, in modern debates it is common to differentiate *phenomenal consciousness*, the subjective 'what it is like' to be in a mental state (Nagel, 1980) with *access consciousness,* understood as the availability of mental states for cognitive functioning (such as

[38] See Lavazza and Robinson (2014) for contemporary defenses of dualism(s) of various sorts.

[39] It should be stressed that, although functionalism is not strictly speaking a physicalist view, it is closely associated with physicalism due to the fact that the sort of states that play the relevant causal roles for the functionalist are (most often) taken to be physical states.

reasoning, action, and verbal report (Block, 1995)). The former captures the 'raw' subjective feel which – recalling Descartes' anti-physicalist meditation – seems particularly difficult to explain in purely physical terms (Chalmers, 1996), whereas access consciousness more readily maps onto cognitive processing and brain function.

In contemporary neuroscience, theories of consciousness have centered on bridging the gap between neural and mental phenomena. Prominent examples include theories that explains this in terms of globally broadcasted information across a neural workspace (Baars, 1997; Dehaene & Naccache, 2001), the cause-effect powers of integrated information (Tononi, 2004), higher-order representations (Rosenthal, 2005), and recurrent processing (Lamme, 2006). A tension that prevails here – partly in virtue of the distinction between access and phenomenal consciousness – is the contrast between theories that emphasize more "thinky" executive prefrontal cognition, such as higher-order theories and global workspace theory, and those that focus on the more perceptual processes associated with regions found in the back-of-the-head, such as integrated information theory and recurrent processing (Seth & Bayne, 2022).

Relatedly, theories of consciousness also differ in terms of whether they emphasize the phenomenology or function of consciousness. In some theories, such as Integrated Information Theory (Tononi, 2004), the phenomenological properties of consciousness are taken as a fundamental, and acts as starting points to identify the processes that have them (subjective experience being intrinsic to any system with integrated information above a certain threshold). In other theories, such as Global Workspace Theory (Baars, 1997; Dehaene & Naccache, 2001), subjective experience is instead viewed as emerging from the functional architecture of consciousness – specifically, when information becomes globally 'broadcasted' to multiple parts of the cognitive system.

Now, although many impressive efforts have been made to computationally model human-like rationality (Gershman et al., 2015; Lewis et al., 2014), autonomy (Ezenkwu & Starkey, 2019), and consciousness (Blum & Blum, 2022; Cleeremans, 2005; Dehaene et al., 2014), they are naturally pestered by the conflicting conceptions that trickle down (and up) from philosophy and neuroscience (Kirkeby-Hinrup et al., 2024).

In a nutshell, different views yield different answers to whether and to what extent machines can be conscious. Under certain assumptions, there appear to be no fundamental obstacles for artificial consciousness. For instance, a prominent specimen of functionalism is *computational* functionalism, according to which consciousness is computational in nature. Just as a computer can run the same program on different hardware, consciousness is seen as patterns of information processing that could potentially be realized in multiple physical substrates. This means, theoretically, that a sufficiently complex computer could be conscious if it had the right functional organization and performed the right kind of computations (conversely, it also means that humans are also, in some important sense, computers). Interestingly, a recent survey conducted by prominent cognitive scientists and philosophers suggests that, if one accepts computational functionalism, there are no obvious technical barriers to constructing AI systems which satisfy indicators for consciousness as defined by several leading theories of consciousness (Butlin et al., 2023).

However, functionalism does not offer a straight-forward bridge across the vast ocean between mind and matter. As I note in Paper VI: "Is a stage magician performing *real magic* if the tricks and illusion manage to fool an audience into believing that the magician has supernatural powers? Is the *illusion of magic* the same as *real magic* if they are functionally equivalent?" (VI, p. 9). And as I get into in Paper VII, a similar dissatisfaction with functionalism (and materialism) can be fueled with additional support from a number of famous arguments in philosophy of mind: by asking whether zombies are logically possible (Chalmers, 1996), whether one could ever feel what it is like to be a bat (Nagel, 1974), or as discussed in Section 2.2, whether Chinese rooms could genuinely understand Chinese (Searle, 1980).

But if one instead adopts a dualist perspective (or some other more restrictive view), the inner mental life might remain forever opaque to all but the experiencing subject. This epistemic barrier would leave us fundamentally uncertain about whether machines are conscious.[40] This impasse could in turn pave the way for one of the two unsettling scenarios. In the first, we can suppose that we – e.g., as a moral community – accept a functionalist view at face value (or some other, less restrictive view), which leads us to recognize robots as conscious beings. Suppose that we also take them to satisfy every relevant functional criteria for phenomenal consciousness along with the relevant capacities for autonomy and rationality to qualify as full moral agents. But *if* functionalism is mistaken, there is a risk that we are fundamentally deceived: *the robots never felt anything on the inside, they just acted as if they did!*[41] But the alternative scenario may be even more repugnant: if we dismiss the possibility of machine consciousness – e.g., due to an anthropocentric attachment to views that would do so – and as a consequence, never welcome

[40] This is referred to as the 'epistemic objection', both in discussions about artificial moral agency and artificial moral status (Behdadi & Munthe, 2020; Dung, 2022; Johansson, 2010). See also Andreotta (2021) for a recent discussion of the hard problem of consciousness in the context of AI rights.

[41] This scenario is vividly explored by Bostrom (2014, p. 173), who describes it as a "Disneyland without children".

artificial entities into our moral communities (as patients or agents), it could lead to an explosion of artificial suffering and exploitation if they in fact *were* conscious.[42]

Challenge 2: Are we really autonomous, free, and rational?

In section 5.2, I tried to establish the intuitive picture that a person's moral agency relies on a capacity for rational and autonomous deliberation, free from external influence and control. However, this picture can be challenged by a growing body of empirical findings in moral psychology suggesting that humans are far from being as autonomous, rational, and free as we might think.

Daniel Wegner's work, for instance, suggests that the sense of voluntary control we experience having over our actions may be more of a retrospective narrative (Wegner, 2004). Another example is Jonathan Haidt, who contends that moral judgements are primarily intuitive gut reactions, with rational deliberations serving mainly as 'post-hoc rationalizations' – we feel first and justify later (Haidt, 2001). The list goes on: experiments on choice blindness by Johansson and colleagues reveal how malleable and unreliable our perceived decision-making processes can be (Johansson et al., 2005), Nisbeth and Wilson's studies on introspection demonstrates that humans have no reliable access into their own mental processes (Nisbett & Wilson, 1977; Wilson, 2004), and the work of Kahneman and colleagues shows how humans, due to various shortcuts and biases, make predictable, systematic errors in judgement and choice (Kahneman, 2011; Kahneman et al., 1982).

Taken together, these findings suggest that our intuitive picture of moral agency might not be a description of capacities that humans actually have, but rather, a kind of philosophical fiction – an idealized projection of what we aspire to be, rather than what we genuinely are. Far from being autonomous rational agents, we appear to be complex biological machineries, mainly driven by unconscious processes that we only partially understand.

This raises serval interesting questions about the prospect of artificial moral agency. If human moral agency is not as robust as traditionally imagined, does this imply that machines – potentially free from biases, cognitive shortcuts, illusions of control, and with reasoning capacities beyond gut feelings – could in some respects be 'fuller' moral agents than us? Alternatively, should we lower the bar for moral agency to reflect a more realistic assessment of human cognition? Or might these findings indicate that our assumption – that we can analyze moral agency in terms of capacities – is misguided?

[42] It is the possibility for such scenarios that makes Metzinger (2021) call for a global moratorium on research that directly aims at developing artificial consciousness. See also Dung (2023) for an 'erring on the side of caution' strategy to deal with this possibility in the face of uncertainty.

Challenge 3: Should we create machines that are rational, autonomous, and conscious?

I am inclined to believe that, in principle, the possibility for creating rational, autonomous, and conscious machines is more compelling than the impossibility. My strongest reason for this belief is that, whatever the true nature of RAC are, the fact that they did emerge in beings like us suggests that they were not the result of some special, non-physical ingredient, but are products of natural processes. This argument, I believe, also holds for more moderate and neutral views. If one holds a moderately dualist view, e.g., acknowledging that, although phenomenal experience may have certain qualities that cannot be reduced to material or functional processes, it is still the result of certain physical and functional processes. There could be something about experiences that can only be understood in phenomenal terms (e.g., an experience of redness as an experience), yet, this is consistent with the idea that *without* the appropriate functions and/or physical processes, there would not be any experiences at all.

With this, however, I am not claiming that it would be a *practically* easy task to reproduce the appropriate conditions and processes that give rise to the mental phenomena of a distinctively human mind (as less restrictive functionalistic views could suggest). Nor am I claiming that current paradigms at the forefront of AI development are pursuing the sort of research trajectories that would take us there.[43]

Regardless, before pursuing the creation of rational, autonomous, and conscious machines, we must confront a critical question: should we do so, and for what reasons?

In section 2.1, I critically discussed five reasons for creating moral machines. And throughout this project, I have repeatedly emphasized one reason in particular: to better understand human morality. But as the project of artificially creating 'full moral agents' (and not merely 'explicit ethical agents') in many ways resemble the project of recreating a 'full' human being, the ethical challenges are both vastly more numerous and significant.

As other controversial scientific endeavors, the drive to create machines with these capacities may reflect a modern form of Promethean hubris – the impulse to challenge divine boundaries. A pertinent analogy can be made to the ethics of cloning, which raises numerous risks and convoluted issues about identity, natural contra unnatural, and social relationships. As a consequence, while there is no

[43] To the contrary, by reading naturalistic philosophers such as Daniel Dennett (1991, 2017), Peter Godfrey-Smith (1996, 2016a, 2016b), Peter Gärdenfors (2024), and Anil Seth (2021), I think we have much to learn about life in the *living* world – and of the neurobiology of consciousness as manifest in through-and-through *living* systems – to identify those trajectories.

single global treaty that universally (and unambiguously) bans all forms of human cloning, the practice is strictly regulated or outright banned across most countries (Langlois, 2017).

I believe that ethical deliberations of a similar magnitude must precede any attempt to create a conscious, autonomous, and rational machine. Two issues stand out in the AI ethics literature. The first has already been discussed: the ethics of artificial consciousness. If some research effort manages to create synthetic consciousness capable of experiencing pleasure and pain – intentionally or unintentionally – there is a risk that it leads to various forms of artificial suffering and exploitation. One drastic solution, as suggested by Thomas Metzinger, is to strictly ban "all research that directly aims at or knowingly risks the emergence of artificial consciousness" (Metzinger, 2021, p. 43). A more moderate alternative is to establish principles and policies for conducting responsible research on AI consciousness. To this end, Butlin and Lappas (2025) have recently proposed that that AI consciousness research should be guided by principles emphasizing harm prevention, adopting a gradual approach with strict safety and risk protocols, promoting transparent knowledge sharing while preventing misuse, and maintaining honest communication about uncertainties and risks.[44]

The second issue, which I have largely avoided in this thesis but remains central to AI ethics, are the various risk related to the creation of superintelligence.[45] For many, this is a concern of an existential kind, fearing that superintelligent artificial intelligence may lead to human extinction. Those who openly warn of an AI apocalypse are often referred to as "AI doomers", and over the last decade, their worries have not only moved from science-fiction to mainstream, but into policy discussions among world leading politicians.[46] What motivates the worry of an AI apocalypse due to the creation of an AGI is the plausible idea that superior cognitive

[44] A similar view is articulated in an open letter from the Association for Mathematical Consciousness Science – signed by leading AI and consciousness researchers – which states: "As AI develops, it is vital for the wider public, societal institutions and governing bodies to know whether and how AI systems can become conscious, to *understand* the implications thereof, and to effectively address the ethical, safety, and societal ramifications associated with artificial general intelligence (AGI)." (AMCS, 2023).

[45] For accessible introductions, see, e.g., Bostrom (2014); Häggström (2016); Russell (2019); Tegmark (2017); Yampolskiy (2024).

[46] For instance, in a single-sentence statement released in May 2023, more than 350 world-leading AI researchers and tech CEO's proclaimed that: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, 2023). Later in the same year, existential risk from artificial general intelligence (AGI) was a major focus on the AI Safety Summit, which was attended by several heads of state (Department for Science, 2023; Fullbrook, 2023).

abilities – e.g., for autonomy and rational thought – also entail a decisive strategic advantage over any other form of intelligence (such as human forms). And given that a superintelligent being may have goals or behaviors that come into conflict with human lives, it yields a host of potential pathways to disaster.

On my own contribution

The challenges outlined above are only three of potentially many more obstacles to establishing whether machines can be full moral agents in virtue of human-like RAC. And as spoiled in section 2.2, this thesis has not succeeded in yielding any conclusive answers. Nor have I managed to resolve any of the three aforementioned challenges that complicate that inquiry. But I hope to have at least shown that it is not only a notoriously tricky terrain, but one which any contribution to the prospect of artificial moral agency necessarily must navigate. I will now describe how various parts of this terrain have been explored across the seven papers.

In Papers II and III, I present a more optimistic vision about the technical and normative prospect of artificial moral agency. Given that it is both (currently) unfeasible and ethically problematic (based on the issues just discussed) to equip artificial agents with human-like RAC, I argue that the development of artificial moral agents should instead be driven by functional capacities that are shaped by normative deliberations on how AI systems should be involved in human practices. Instead of full-blown RAC-capacities, I suggest that it is possible – though by no means necessary – to model *rationality* in a way that "allows artificial agents to effectively pursue goals without necessarily relying on the meta-cognitive abilities of human rationality"; *autonomy* in ways that "enable human operators to oversee, intervene, or share the control of the system to avoid unwanted consequences"; and *consciousness* in terms of reward functions in learning systems that "functionally mimic aspects of the role subjective preferences have in human cognition, without the phenomenological experience of suffering" (Paper II, p. 7).

Here, one can reasonably object to whether my functionalistic interpretation of virtue ethics in the context of reinforcement learning – such as virtues and vices, eudaimonia ('flourishing'), and phronesis ('practical wisdom') – bears any relevant resemblance to the human qualities that inspired them. I acknowledge that for some, it may seem extremely provocative to even propose that such deeply human aspects of moral life could ever be given a computational characterization.[47] But such criticisms would miss the point. The framework is not meant to give a recipe for creating artificial humans. Rather, it intended as a generic and adaptable blueprint that offers a wide range of options for constructing agents capable of serving

[47] Relatedly, one could also note that the presented framework is based on a collection of assumptions and views that – like computational functionalism in challenge 1 above – make the prospect of artificial virtuous agents seem both plausible and desirable, yet, object to whether these assumptions and views are themselves sound.

morally significant roles in human practices, given that it *is* desirable for them take on such roles. As such, it is not a final product but a first step in a longer discussion on how to leverage insights – from human moral cognition and virtue ethics – in the development of computational systems imbued with human-centered notions of morality.

In Paper I, I analyze how RAC have been defined and interpreted in philosophy and computer science. Here, I take more of a meta-perspective, arguing that the conceptual discrepancies that emerge lead to problematic forms of interdisciplinary confusion, which in turn can solidify incommensurable perspectives regarding the near- and long-term challenges of AI. One example is how the term 'autonomy' in the Kantian tradition entails a form of self-legalization – i.e., acting according to one's self-formulated rules – whereas in AI development it most often refers to the ability to perform a certain task independent of human supervision or control.

In Paper IV, however, I present a more skeptical view, arguing that machines – both now and in the foreseeable future – lack the capacities needed to justify normative theory as a 'top-down' strategy for implementing ethics into machines; and it is perhaps no coincidence that the capacities I discuss are closely tied to RAC.[48]

In Paper V (section 5.2.3, pp. 51-56), I explore other complications, arguing that even if we accept relatively liberal views on consciousness – in particular, that it is computational in nature – there are a range of additional problems associated with any system's ability to understand and process the semantics of moral language. More generally, the paper can also be seen as substantial contribution to challenge 2 above – on how ethical decision-making are constrained and enabled with respect to available resources, heuristics, and cognitive capacities. The most significant example is the Moral Tractability Thesis (section 7.3, pp. 72-73), which I argue can serve to specify the limits of ethical decision-making in light of bounded rationality.

Paper VI uses Ishiguro's *Klara and the Sun* (Ishiguro, 2021) as vehicle to explore the three challenges described above. It analyses RAC by examining their role in moral agency against a backdrop of related ideas: the Aristotelian notion of humans as rational animals, the immutable soul central to Abrahamitic traditions, and the autonomy and freedom of citizens in liberal democracies. In the context of challenge 2, this led me to reconsider RAC, not as innate capacities defining what humans are, but as cultural constructs of the idealized qualities we hope to have. Over millennia, these ideals have been cultivated through specific socio-political movements,

[48] However, it should be stressed that my main argument in Paper IV – to shift from theory-first to practice-first – can be made independently of whether machines can or cannot have RAC. The point is not to argue that machines would only be moral agents given that they were rational, autonomous, and conscious; rather, the point is that it is problematic to start the inquiry from that direction.

eventually becoming entrenched as fundamental aspects of (a predominantly Western) self-understanding.

I believe the most valuable contribution from Paper VI, however, resides not so much in what it argues as in how vividly it illustrates the intricate subtleties of the ethical issues we may face in a possible near future with artificial companions (which should mainly be credited to Ishiguro's literary genius). We follow an artificial being – Klara – from the "inside": a being that, according to some theories, should qualify as possessing both moral status and agency, yet for a variety of reasons is not (consistently) recognized as such. This provides an empathetic insight into what it is like to be seen and treated by others in light of the plethora of views (e.g., on consciousness and personal identity) that circulate in human societies, as embodied by the characters of the novel: the sentimental mother (attached to the idea that there is something special inside us), the functionalistic roboticist (Capaldi), the ambivalent neighbor (Helen), and the close friend (Josie).

By the time I wrote Paper VII, which in some sense provides a condensed summary of the above journey, my views had been saturated by the realization that artificial RAC depends on a number of contested answers to what may be fundamentally unsolvable questions. Instead, this led me to use the insight from Paper VI to explore moral agency, not *only* in terms of capacities, but also with respect to how those capacities are situated within our moral practices and relationships. And it is this that we will turn to next.

5.4. Moral agency in practice

It may be a natural starting point to think of a moral agent as someone *capable* of reasoning and acting based on what is right and wrong, and thus, to analyze moral agency in terms of this capacity. Most often, however, we do not seem to care too much about capacities as such. Say, if you did something to me that I found morally wrong, I might say, "hey, why did you do that?", and then inquire whether you are the sort of morally sensitive person one could expect *not to commit* that transgression. Having concluded – yes, you are that sort of person – I might blame you for your actions ("you should not have done that!"). Yet, before the transgression occurred, I would not dwell on whether you had the capacity *not to do it*. We do not, at least not typically, walk around with checklists of criteria others must satisfy to qualify as moral agents in a way that is independent of our interactions. Rather, the relevance of capacities only comes into play when they *are* actualized in social interactions – how our actions respond to and are responded to by others in a shared social context.

These social contexts vary in innumerable ways. Consider, for instance, how contexts like families, tribes, neighborhoods, workplaces, societies, nations, and the global community differ spatially and temporally, spanning vast or confined spaces and long or short timeframes, and in terms of interaction intensity, frequency, and stability. Some people we meet daily over years; others, we encounter once.

Social contexts also differ in terms of their normative elements. Think of a close friendship between two people. Such a relationship may have a large set of morally relevant elements that are unique to that particular relationship: of love, care, shared memories, and a complex baggage of expectations. It is, for instance, common to expect those closest to us to adhere to specific moral standards (e.g., treating us in ways we prefer). But it would be unreasonable – and frankly, quite odd – to expect any random stranger to adhere to the same standards. A similar observation can be made by considering how social norms differ across the world, where behaviors that are normal or even encouraged in some cultures are frowned upon in another.

This suggests an alternative approach to defining a moral agent: someone who is responsive to the normative features – e.g., expectations, demands, obligations, rules – that prevail in a specific social context. In this approach, which I call 'practice-first', we start from a moral practice, which may *then* allow us to formulate a checklist of capacities required to qualify as a moral within that practice. This reverses the direction found in the 'capacities-first' approach discussed in the previous section, where we start from capacities and *then* assess whether someone has them. This reversal shifts the focus from a more abstract inquiry into the universal and objective features of moral agency (e.g., RAC), to a more concrete exploration of *particular* forms of moral agency that are relative to social contexts.

In Paper VII, I discuss various benefits and drawbacks of both approaches. For instance, a capacity-first view, if sound, can help determine whether some class of beings – say, teenagers, a non-human animal species, or robots – should be regarded as a moral agent based on descriptive features of their capacities in a way that transcends the particularities of a certain context. It is not just 'whatever goes on in the practice' but an inquiry that can lead to clearer and more generalizable criteria to coordinate our moral practices on a larger scale. If it succeeds, it allows us to make inferences of the form: (P1) all entities that have XYZ (e.g., RAC) are moral agents, (P2) entity E (or class of entities) has XYZ, (C) and thus conclude that E is a moral agent. On the other hand, by centering on concrete social contexts, practice-first views seem to better account for the rich nuances of what being together entails in different communities, cultures, and political arrangements.

These approaches are not necessarily mutually exclusive. Problematically, however, is that they sometimes yield diverging answers to questions about who is or should be regarded a moral agent. Artificial beings present such a problematic case. Consider, for example, sophisticated AI companions. Such companions might be regarded as moral agents based on external practices, such as how humans engage with them, even if they potentially lack the relevant internal capacities (e.g., RAC). Conversely, there are examples of beings who possessed the internal capacities for moral agency yet were (or are) systematically denied recognition as such by external practices. Consider instances of slavery or misogyny, where individuals were not recognized as moral agents by societal or legal systems, despite their evident capacity for participating in such practices.

In Paper VII, I describe how this tension – between 'internal' capacities and 'external' features of moral practices – is reflected in two different debates. The first is the conflict between *being* responsible and practices of *holding* responsible as competing grounds for responsible agency, in the debate following Strawson's *Freedom and Resentment* (Strawson, 1962). According to one camp (typically referred to as 'Strawsonians'), responsibility is, in some way, grounded in our practices of holding others responsible (Shoemaker, 2017; Wallace, 1994; Watson, 1993), whereas others maintain that justifiably holding someone responsible presupposes their being responsible (Brink & Nelkin, 2013; Fischer & Ravizza, 1993).[49]

The second is the debate about the moral status of artificial beings, such as social robots. According to one camp, the attribution of moral status depends on intrinsic properties, such as the capacity to experience pleasure and pain (Andreotta, 2021; Dung, 2022; Mosakas, 2021), whereas 'relational' approaches hold that moral status does not depend on intrinsic properties, but is attributed in virtue of social relationships (Coeckelbergh, 2010, 2014; Gunkel, 2014, 2018). Typically, the latter view is more accepting of the idea that AI systems can have some form of moral status, whereas the former view is less so.

Instead of advocating over one over the other, Paper VII presents a view in which the two sides – the 'inner' and 'outer' – are co-constructive aspects of the same unified whole. To support the central idea, I provide examples of how moral capacities and moral practices are co-constructed across multiple timescales and social arrangements – natural and cultural history, individual development, and everyday interactions – which are continuously stabilized and revised via various mechanisms. I then elaborate on three broader appeals of the view: (i) that it gives a theoretical coherence not afforded by either capacities- or practice-oriented accounts, (ii) that it can preserve the strengths and mitigate the drawbacks inherent to each view, (iii) and how it sheds new light on ontological, epistemological, and metanormative aspects of moral agency, roughly corresponding to what the *nature*

[49] See also the recent dissertation by Dorna Behdadi (2023), which presents a practice-oriented approach to the moral agency of nonhuman animals and artificial entities.

of moral agency, how we get to *know* that nature (and whether others have it), and what it *should* be.

Finally, I also elaborate on what the view could say about the moral agency of artificial agents. In short, I argue that although human-machine interactions have not *yet* achieved the sort of stability and regularity required to generate more substantial and generalizable normative features – as characteristic of human-human interactions – the view shows what it would take for such interactions to one day become potent sources for reason-guiding with respect to larger social contexts.

However, I also note that this line of reasoning assumes that human-human interactions should serve as the benchmark for moral agency. As an alternative, I also discuss another, arguably more exciting opportunity: to let machines realize completely *new* forms of relationships, that need not – and arguably should not – use human-human relationships as a metric.

Chapter 6 – For those who might make it to the robot lab

In this thesis, we have looked into various aspects of how to build nice robots. There was a hope to one day reach a robot lab where we could build these nice robots. Unfortunately, we are not there just yet. Instead, we have taken a detour into questions that should be addressed before we can confidently say that we have found the right blueprint. Fortunately, there is still hope for others to make it, especially if they follow these two sets of recommendations, all of which are based on lessons described in this thesis.

6.1. Thesis summary and conclusion

Moral machines

The first set of recommendations comprises the lessons covered in Chapters 2-4 and Papers I-V. Let's begin with the challenges. In Chapter 2, we saw how important it is to justify our moral machine building ventures with compelling reasons, and to carefully address the various risks and drawbacks that such ventures could entail. Additionally, in Chapter 3, we saw how critical it is to account for the fact that morality is not only a multifaceted and heterarchical phenomena that we deeply disagree about, but that it is also hard. An additional set of challenges emerged in Chapter 4, regarding the vast space of particularities that need to be considered in the development of moral machines, the critical role of interpretation throughout all stages of development, how small details can matter as much as the bigger picture, and the impossibility of comparing – or benchmarking – the moral performance of machines.

To address these challenges, it was recommended for machine building endeavors to pursue technical and normative aspects in unison (2.1), be attentive to how normative theory serves human practices (2.2), be informed by insights, methods, and theories from multiple disciplines (3.1), see disagreements as opportunities rather than obstacles (3.2), and use unifying frameworks as common ground (4.3.5). Furthermore, it was demonstrated how we can learn about ethical decision-making for both humans and machines by exploring how it is enabled and constrained by computational resources (3.3 & 4.3.3). Finally, instead of implementing normative theories in a 'top-down' fashion – from theory to practice, it was recommended to start from the other way around (4.3.2), with a careful attention to the practice and how normative theories converge in their prescriptions (4.3.1).

Moral agency

The second set of recommendations comprises Chapter 5 and Papers VI-VII (and to a less extent, Papers I-V). This concerned questions about whether machines can be something more than just machines – in particular, if they can qualify as 'full moral agents', with all the features of moral agency we typically associate with human adults. First, this exploration centered on *capacities* for moral agency, focusing on rationality, autonomy, and consciousness (5.2 & 5.3). Given the deep uncertainties, conceptual confusion (Paper I), and ethical concerns (Paper VI) surrounding this question, it was argued that artificial systems should not aim to replicate human moral agency, but rather, be designed to align with the normative features of the human practice in which they operate (5.3 and Papers II, III, IV & VI). It was also argued that analyzing moral agency exclusively in terms of capacities is insufficient, which motivated an inquiry into how moral agency is situated in relation to social practices (5.4 & Paper VI-VII). In an attempt to reconcile a split between moral capacities and moral practices, it was argued that they should be viewed as co-constructive aspects of the same unified whole (Paper VII).

Conclusion

As I mentioned in the introduction, this work is not intended to present a coherent vision or argument, but rather to serve as a recipe book. As anyone who has followed a recipe can attest, certain things should be taken with a pinch of salt. With that said, it is possible to distill two main threads of this book of recipes: one practical and one theoretical.

From practice to machine implementation – The first thread concerns how machines with capacities for moral decisions and actions should be developed for concrete, practical contexts. For this project, I have presented several contributions: overcoming interdisciplinary boundaries between moral philosophy and computer science (Paper I); demonstrating how virtue ethics can leverage insights from human moral cognition in ways that resonate with modern machine learning methods (Papers II & III); and situating normative ethics to inform machine design in a way that better captures moral practice (Paper IV). Additionally, I have analyzed how ethical decision-making can be realized via algorithmic methods with distinct trade-offs – such as rule-following, causal reasoning, and learning – and how it is bounded by implementation-invariant resources (such as time and memory) while being complicated by issues related to inductive biases, strategic dynamics, and inexplicability (Paper V). If a moral machine-building venture proceeded in this 'bottom-up' fashion, with careful attention to both moral and non-moral aspects, I believe that it not only can be compellingly justified, but lead to the creation of robots that we can call 'nice'.

From machine implementation to theory – The second thread is not about building moral machines, but rather, about how approaching morality from a computational perspective offers exciting opportunities to integrate previously disconnected interdisciplinary insights, ultimately contributing to new understandings of morality itself. This, I hope to have demonstrated in my work virtue ethics (Papers II & III), game theory (Papers II & V), and computational complexity (Paper V). In this thread, the important thing is not about getting to lab, but to sit back and ask: why do we think and act like we do, given the intricacies of our biological nature, psychology, and social world? And why should we treat each-other in certain ways rather than others? As we approach these questions through the lens of a moral machine builder, we discover new avenues to answer these questions – avenues that might otherwise remain hidden.

Interestingly – as I wrote in the introduction regarding the terms 'robots', 'nice', and 'build' – these threads demonstrate two additional ways in which the title of this thesis is somewhat misleading. Yet, I hope they can still serve as a guide, both for those who wish to get to the lab, and for those who prefer to sit back and ask more questions.

6.2. Exposition of papers

Paper I – Interdisciplinary Confusion and Resolution in the Context of Moral Machines (Stenseke, 2022a) – seeks to resolve the conflicts and confusion that arise from *building* and *thinking about* moral machines, and describes how fruitful synergies can be achieved from doing both. In particular, it explores the tension between discipline-specific approaches to moral machines, and presents both practical and theoretical ways to alleviate those issues in order to foster inter- and transdisciplinary research in the field of machine ethics. The paper takes its starting-point in two prevalent approaches to machine morality: (i) *the philosophical approach to machine ethics* (PME), which centers on conceptual exploration of what computational systems *ought* to do (and correspondingly, what systems *ought* to be built); (ii) *the engineering approach to machine ethics* (EME), which centers on exploring the kind of morality that *can* be implemented in computer systems (and what moral systems *can* be built). The paper then describes how practices, concepts, and aims inherent to these disciplinary perspectives may facilitate incommensurable views on the prospect of machine morality, which in turn curtails what one discipline could meaningfully contribute to the overarching challenges of the field. As a remedy, the paper presents several ways to avoid the aforementioned issues and promote fruitful synergies for interdisciplinary collaborations in machine ethics, focusing on the strengths of each perspective, and how to avoid their pitfalls. Finally, the paper describes how *metacognitive scaffolds* can be used to articulate disciplinary perspectives, effectively providing means to foster communication and resolve epistemic and normative conflicts in interdisciplinary research projects.

Paper II – Artificial Virtuous Agents: From Theory to Machine Implementation (Stenseke, 2023a) – presents and argues for *virtue ethics* as a recipe for the construction of moral machines, and describes how the theory can be taken all the way from theory to machine implementation. The paper begins by discussing four major appeals and four major challenges for computational approaches to virtue ethics. It then outlines a path to artificial virtuous agents based on moral functionalism, bottom-up learning, and eudaimonic reward, which is translated to a generic cognitive architecture for computational implementation.

Paper III – Artificial Virtuous Agents in a Multi-agent Tragedy of the Commons (Stenseke, 2024a) – expands the work of paper II in two ways: (i) it demonstrates the promise of artificial virtue ethics in a simulation with game-theoretic dilemmas (called *BridgeWorld*), and (ii) digs deeper into some of the remaining technical and philosophical challenges for artificial virtue ethics.

Paper IV – The Use and Abuse of Normative Ethics for Moral Machines (Stenseke, 2023b) – critically examines the methodological strategy of using normative theories as blueprints for the construction moral machines, arguing that machines currently lack many of the resources that are needed to justify the very use of normative theory.

Paper V – On the Computational Complexity of Ethics: Moral Tractability for Minds and Machines (Stenseke, 2024b) – uses *computational complexity* to analyze what kind of moral machines are possible based on what computational systems can or cannot do with bounded computational resources (e.g., time, knowledge, learning). It is demonstrated that nearly all problems that prevalent normative frameworks pose – consequentialism, deontology, and virtue ethics – lead to intractability issues. The paper also provides several insights about the computational nature of normative ethics and discusses how computational complexity have implications for both philosophical and cognitive-psychological research on morality by advancing the Moral Tractability Thesis.

Paper VI – The Morality of Artificial Friends in Ishiguro's *Klara and the Sun* (Stenseke, 2022b) – explores whether artificial entities can have a moral status or be moral agents on the basis of Ishiguro's novel *Klara and the Sun* (2021). It juxtaposes two approaches to these questions, the view “from within” and the view “from outside”, and argues that the book exposes the shortcomings of the first in relation to the social reality imposed by the second. That is, regardless of what moral qualifiers one can attribute to artificial beings such as Klara “from within”, they are ultimately determined by the views of others (“from outside”). The interesting upshot is that others’ view not only include typical features of the view “from outside” (e.g., behaviors and functions), but can also involve metaphysical features about the nature of consciousness and personhood.

Paper VII – Knowing and Owing Each-Other: On the Co-construction of Moral Agency Across Time and Space (Stenseke, Unpublished manuscript) – expands upon the ideas of paper VI in an attempt to present a new way of thinking about moral agency. The paper takes it starting point in a tension between two approaches to moral agency. The first centers on *capacities* that a being must have in order to qualify as a moral agent. The second centers on how agents behave and relate to each-other in virtue of a specific moral *practice*. While the two approaches tend to converge for communities of humans, they lead to seemingly unresolvable meta-theoretical puzzles when we consider non-human animals, artificial entities, and other hard cases. As an alternative, the paper describes a way of thinking about moral agency in which the two approaches are co-constructive sides of the same whole, where (internal) moral capacities and (external) moral behaviors act to either stabilize or revise what it is to be moral agent with respect to some social context. I explain how this dynamical interplay works across multiple timescales – from natural history to day-to-day interactions – and social arrangements via a variety of mechanisms. The emerging view, I argue, yields a holistic picture of moral agency that is in a better position to accommodate the strengths and drawbacks of capacity-oriented and practice-based views. Finally, I expand on how the co-constructive view illuminates the prospect of artificial moral agency.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement Learning as a Framework for Ethical Decision Making. *AAAI Workshop: AI, Ethics, and Society*, *16*(2).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, *7*(3), 149-155. https://doi.org/10.1007/s10676-006-0004-4

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility. *Robot ethics: The ethical and social implications of robotics*, 55-68.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12-17.

AMCS. (2023). *Association for Mathematical Consciousness Science. The Responsible Development of AI Agenda Needs to Include Consciousness Research. Open letter.* Retrieved 6 March 2025 from https://amcs-community.org/open.letters/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI magazine*, *28*(4), 15-15.

Anderson, M., & Anderson, S. L. (2008). ETHEL: Toward a Principled Ethical Eldercare System. *AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems*, *2*.

Anderson, M., & Anderson, S. L. (2010). Robot be good. *Scientific American*, *303*(4), 72-77.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Andreotta, A. J. (2021). The hard problem of AI rights. *AI & SOCIETY*, *36*(1), 19-32.

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Arkin, R. C. (2007). *Governing lethal behavior: embedding ethics in a hybrid deliberative/hybrid robot architecture*.

Armstrong, S. (2015). Motivated value selection for artificial agents. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence,

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics & Philosophy*, *16*(2), 247-266.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of political economy*, *58*(4), 328-346.

Asimov, I. (1950). Runaround. In *I, Robot*. Doubleday.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M., Everett, J. A., Evgeniou, T., Gopnik, A., & Jamison, J. C. (2022). Computational ethics. *Trends in cognitive sciences*, *26*(5), 388-405.

Axelrad, S. (1951). *Field theory in social science: selected theoretical papers by Kurt Lewin*. Taylor & Francis.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390-1396.

Ayer, A. J. (1936). *Language, truth and logic* (Vol. 47). V. Gollancz.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., & McKinnon, C. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Behdadi, D. (2023). *Nonhuman Moral Agency: A Practice-Focused Exploration of Moral Agency in Nonhuman Animals and Artificial Intelligence* [Doctoral Thesis (monograph)]. Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*, 195-218.

Bentham, J. (1780). *An Introduction to the Principles of Morals and Legislation* (Vol. 45). Dover Publications.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/https://doi.org/10.1016/j.cognition.2018.08.003

Blackburn, S. (1993). *Essays in quasi-realism* (Vol. 1). Oxford University Press.

Blackburn, S. (2003). *Ethics: A very short introduction* (Vol. 80). Oxford University Press, USA.

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, *18*(2), 227-247.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences*, *119*(21), e2115934119. https://doi.org/doi:10.1073/pnas.2115934119

Boddington, P. (2023). AI Ethics. *Singapur: Springer International Publishing*, 48.

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

Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. *Robot ethics: The ethical and social implications of robotics*, 85-108.

Brink, D. O., & Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. *Oxford studies in agency and responsibility*, *1*, 284-313.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Butler, S. (1863). Darwin among the Machines. *The Press*.

Butlin, P., & Lappas, T. (2025). Principles for Responsible AI Consciousness Research. *arXiv preprint arXiv:2501.07290*.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., & Ji, X. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Cardwell, D. S. L. (2001). *Wheels, clocks, and rockets: A history of technology*. WW Norton & Company.

Center for AI Safety. (2023). https://www.safe.ai/work/statement-on-ai-risk

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501-532.

Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation*, *8*(2), 278-296.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Champagne, M., & Tonkens, R. (2015). Bridging the Responsibility Gap in Automated Warfare. *Philosophy & Technology*, *28*(1), 125-137. https://doi.org/10.1007/s13347-013-0138-3

Cleeremans, A. (2005). Computational correlates of consciousness. In S. Laureys (Ed.), *Progress in Brain Research* (Vol. 150, pp. 81-98). Elsevier. https://doi.org/https://doi.org/10.1016/S0079-6123(05)50007-4

Cloos, C. (2005). The Utilibot project: An autonomous mobile robot based on utilitarianism. *2005 AAAI Fall Symposium on Machine Ethics*, 38-45.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209-221.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology*, *27*, 61-77.

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Crisp, R. (2014). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76-84. https://doi.org/https://doi.org/10.1016/j.conb.2013.12.005

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1-37.

Dehghani, M., Tomai, E., Forbus, K. D., & Klenk, M. (2008). An Integrated Reasoning Approach to Moral Decision-Making. *AAAI*, 1280-1286.

Dennett, D. C. (1991). *Consciousness explained*. Penguin UK.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company.

Department for Science, Innovation and Technology. (2023). *AI Safety Summit 2023*. Retrieved 6 March 2025 from https://www.gov.uk/government/topical-events/ai-safety-summit-2023

Descartes, R. (1637). *Discourse on the Method of Rightly Conducting One's Reason and of Seeking Truth in the Sciences*. Sutherland and Knox (1850).

Dietrich, E. (2001). Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 323-328.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Springer.

Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2024). Large Language Models as Moral Experts? GPT-4o Outperforms Expert Ethicist in Providing Moral Guidance. *PsyArXiv Preprints*, *28*.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in cognitive sciences*, *27*(7), 597-600.

Dung, L. (2022). Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status is Flawed. *Science and Engineering Ethics*, *28*(6), 1-15.

Dung, L. (2023). How to deal with risks of AI suffering. *Inquiry*, 1-29.

Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *arXiv preprint arXiv:1901.00064*.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1-19.

Emilsson, A. (2024). *Minimal Optimism: Reading PF Strawson on Responsibility* [Doctoral Thesis (monograph)]. Department of Philosophy, Lund University.

European Commission. (2024). *AI Act enters into force*. Retrieved 6 March 2025 from https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

Ezenkwu, C. P., & Starkey, A. (2019). Machine Autonomy: Definition, Approaches, Challenges and Research Gaps. In K. Arai, R. Bhatia, & S. Kapoor, *Intelligent Computing* Cham.

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, *49*(3), 363-375.

FeldmanHall, O., & Mobbs, D. (2015). A neural network for moral decision making. In M. L. AW Toga (Ed.), *Brain Mapping: An Encyclopedic Reference.* Elsevier.

Fischer, J. M., & Ravizza, M. (1993). *Perspectives on moral responsibility*. Cornell University Press.

Frank, L., & Klincewicz, M. (2016). Metaethics in context of engineering ethical and moral systems. *2016 AAAI Spring Symposium Series*.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, 55-95.

Fullbrook, D. (2023). *AI summit brings Elon Musk and world leaders to Bletchley Park*. BBC News. Retrieved 6 March 2025 from https://www.bbc.com/news/uk-england-beds-bucks-herts-67273099

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411-437.

Gauthier, D. (1986). *Morals by agreement*. OUP Oxford.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278. https://doi.org/doi:10.1126/science.aac6076

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press.

Gips, J. (1994). Toward the ethical robot. In K. M. Ford, C. N. Glymour, & P. J. Hayes (Eds.), *Android Epistemology* (pp. 243--252). MIT Press.

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Godfrey-Smith, P. (2016a). Mind, Matter, and Metabolism. *Journal of Philosophy*, *113*(10), 481-506.

Godfrey-Smith, P. (2016b). *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.

Govindarajulu, N. S., Bringsjord, S., Ghosh, R., & Sarathy, V. (2019). Toward the engineering of virtuous machines. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 29-35.

Greaves, H. (2017). Population axiology. *Philosophy Compass*, *12*(11), e12442.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108.

Gunkel, D. J. (2014). A Vindication of the Rights of Machines. *Philosophy & Technology*, *27*(1), 113-132. https://doi.org/10.1007/s13347-013-0121-z

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gärdenfors, P. (2024). *Kan AI tänka? Om människor, djur och robotar*. Fri Tanke.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833-838. https://doi.org/10.1038/s43588-023-00527-x

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, *7*(1), 17-52.

Harsanyi, J. (1977). Morality and the Theory of Rational Behavior. *Social Research: An International Quarterly*, *44*(4), 623-656.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*(1), 19-29.

Hobbes, T. (1651). *Leviathan*. Oxford University Press (1996).

Hoffmann, A. A., & Willi, Y. (2008). Detecting genetic responses to environmental change. *Nature Reviews Genetics*, *9*(6), 421-432. https://doi.org/10.1038/nrg2339

Horgan, T., & Timmons, M. (2000). Nondescriptivist cognitivism: Framework for a new metaethic. *Philosophical Papers*, *29*(2), 121-153.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799-819.

Hudson, N. C., Pauloski, J. G., Baughman, M., Kamatar, A., Sakarvadia, M., Ward, L., Chard, R., Bauer, A., Levental, M., & Wang, W. (2023). Trillion parameter ai serving infrastructure for scientific discovery: A survey and vision. *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies*, 1-10.

Hume, D. (1739). *A Treatise of Human Nature: A Critical Edition (2007)*. New York: Penguin.

Hummert, S., Bohl, K., Basanta, D., Deutsch, A., Werner, S., Theißen, G., Schroeter, A., & Schuster, S. (2014). Evolutionary game theory: cells as players. *Molecular BioSystems*, *10*(12), 3044-3065.

Hursthouse, R. (1999). *On virtue ethics*. OUP Oxford.

Hursthouse, R., & Pettigrove, G. (2023). Virtue Ethics. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Vol. Fall 2023).

Häggström, O. (2016). *Here be dragons: Science, technology and the future of humanity*. Oxford University Press.

Ishiguro, K. (2021). *Klara and the Sun*. Faber and Faber.

Jaworska, A., & Tannenbaum, J. (2023). The Grounds of Moral Status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Vol. Spring 2023).

Johansson, L. (2010). The Functional Morality of Robots. *International Journal of Technoethics (IJT)*, *1*(4), 65-73. https://doi.org/10.4018/jte.2010100105

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116-119.

Johnson, M. (2012). There is no moral faculty. *Philosophical Psychology*, *25*(3), 409-432.

Johnson, R. N., & Cureton, A. (2024). Kant's Moral Philosophy. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Vol. Fall 2024).

Joyce, R. (2022). Moral Anti-Realism. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy* (Vol. Winter 2022).

Kagan, S. (2019). *How to count animals, more or less*. Oxford University Press.

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

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kant, I. (1785). *Groundwork for the Metaphysics of Morals*. Yale University Press (2008).

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230254.

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293-307.

Kirkeby-Hinrup, A., Stenseke, J., & Overgaard, M. S. (2024). Evaluating the explanatory power of the Conscious Turing Machine. *Consciousness and Cognition*, *124*, 103736. https://doi.org/https://doi.org/10.1016/j.concog.2024.103736

Kitcher, P. (2011). *The ethical project*. Harvard University Press.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into practice*, *16*(2), 53-59.

Korsgaard, C. M. (1996). *The sources of normativity* (Vol. 110). Cambridge University Press.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, *10*(11), 494-501.

Langlois, A. (2017). The global governance of human cloning: the case of UNESCO. *Palgrave Communications*, *3*(1), 17019. https://doi.org/10.1057/palcomms.2017.19

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2).

Lavazza, A., & Robinson, H. (2014). *Contemporary dualism: A defense*. Routledge.

Law, J. M. (1997). *Puppets of Nostalgia*. Princeton University Press. https://doi.org/doi:10.1515/9781400872954

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational Rationality: Linking Mechanism and Behavior Through Bounded Utility Maximization. *Topics in cognitive science*, *6*(2), 279-311. https://doi.org/https://doi.org/10.1111/tops.12086

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., & Kumar, A. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Malle, B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243-256.

Malle, B. F., & Scheutz, M. (2020). Moral competence in social robots. In *Machine ethics and robot ethics* (pp. 225-230). Routledge.

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots* (pp. 3-17). Springer.

Marr, D. (1982). *Vision: A Computational Approach*. Freeman & Co.

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(01), 43-66.

Mill, J. S. (1863). *Utilitarianism*. Cambridge University Press.

Mirzaeighazi, S. (2024). *On Responsibility and Punishment* [Doctoral Thesis (monograph)]. Department of Philosophy, Lund University.

Mirzaeighazi, S., & Stenseke, J. (2024). Responsibility Before Freedom: closing the responsibility gaps for autonomous machines. *AI and Ethics*, 1-13.

Mishra, A. (2023). Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902

Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., & Bauer, M. (2024). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, *224*(2), 33-35.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18-21.

Moore, G. E. (1903). *Principia ethica*. Dover Publications.

Mosakas, K. (2021). On the moral status of social robots: considering the consciousness criterion. *AI & SOCIETY*, *36*(2), 429-443.

Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159-168.

Nagel, T. (1980). What is it like to be a bat? In *The Language and Thought Series* (pp. 159-168). Harvard University Press.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48-49.

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560-1563.

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275-1289. https://doi.org/10.1007/s10677-016-9745-2

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Parfit, D. (2011). *On What Matters: Two-Volume Set*. Oxford University Press.

Pereira, L. M., & Lopes, A. B. (2020). *Machine ethics: from machine morals to the machinery of morality*. Springer.

Pereira, L. M., & Saptawijaya, A. (2007). Modelling morality with prospective logic. *Portuguese conference on artificial intelligence*, 99-111.

Perrault, R., & Clark, J. (2024). Artificial Intelligence Index Report 2024.

Ridge, M., & McKeever, S. (2023). Moral Particularism and Moral Generalism. *The Stanford Encyclopedia of Philosophy*, *Summer 2023*.

Rosenthal, D. (2005). *Consciousness and mind*. Clarendon Press.

Rousseau, J.-J. (1755). *Discours sur l'origine et les fondemons de l'egalité parmis les hommes*. Marc Michele Rey.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach, 4th ed.*

Scheutz, M. (2016). The Need for Moral Competency in Autonomous Agent Architectures. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 517-527). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_30

Schulte, P. M. (2014). What is environmental stress? Insights from fish living in a variable environment. *Journal of Experimental Biology*, *217*(1), 23-34.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-424.

Sen, A. (1979). Utilitarianism and welfarism. *Journal of Philosophy*, *76*(9), 463-489.

Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), 439-452.

Shim, J., Arkin, R., & Pettinatti, M. (2017). An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation. *2017 IEEE International conference on robotics and automation (ICRA)*, 2936-2942.

Shoemaker, D. (2017). Response-dependent responsibility; or, a funny thing happened on the way to blame. *Philosophical Review*, *126*(4), 481-527.

Simmons, G. (2022). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

Singer, P. (2011). *Practical ethics*. Cambridge university press.

Smart, R. N. (1958). Negative utilitarianism. *Mind*, *67*(268), 542-543.

Smith, M. (1994). *The moral problem*. Blackwell.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77.

Stenseke, J. (2021). Persistent homology and the shape of evolutionary games. *Journal of theoretical biology*, *531*, 110903. https://doi.org/https://doi.org/10.1016/j.jtbi.2021.110903

Stenseke, J. (2022a). Interdisciplinary confusion and resolution in the context of moral machines. *Science and Engineering Ethics*, *28*(3), 24.

Stenseke, J. (2022b). The Morality of Artificial Friends in Ishiguro's Klara and the Sun. *Journal of Science Fiction and Philosophy*, *5*.

Stenseke, J. (2023a). Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*, *38*(4), 1301-1320.

Stenseke, J. (2023b). The use and abuse of normative ethics for moral machines. In *Social Robots in Social Institutions* (pp. 155-164). IOS Press.

Stenseke, J. (2024a). Artificial virtuous agents in a multi-agent tragedy of the commons. *AI & SOCIETY*, *39*(3), 855-872. https://doi.org/10.1007/s00146-022-01569-x

Stenseke, J. (2024b). On the computational complexity of ethics: moral tractability for minds and machines. *Artificial Intelligence Review*, *57*(4), 105. https://doi.org/10.1007/s10462-024-10732-3

Sterkenburg, T. F., & Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese*, *199*(3), 9979-10015.

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187-211.

Søgaard, A. (2023). Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, *33*(1), 33-54.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2016). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, *18*(6), 1429-1439.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, *53*(6), 1-38.

Tomasello, M. (2016). *A natural history of human morality*. Harvard University Press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Truitt, E. R. (2015). *Medieval robots: Mechanism, magic, nature, and art*. University of Pennsylvania Press.

Usher, A. P. (1954). *A history of mechanical inventions*. Courier Corporation.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939-984.

Van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213-218.

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, *25*(3), 719-735.

Velichkov, A. (2023). *Responsibility and Ambivalence* [Doctoral Thesis (monograph)]. Department of Philosophy, Lund University.

Vishwanath, A., Bøhn, E. D., Granmo, O.-C., Maree, C., & Omlin, C. (2023). Towards artificial virtuous agents: games, dilemmas and machine learning. *AI and Ethics*, *3*(3), 663-672.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & SOCIETY*, *22*, 463-475.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Watson, G. (1993). 4. Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 119-148). Cornell University Press.

Way, M. (2017). ”What I cannot create, I do not understand”. *Journal of Cell Science*, *130*(18), 2941-2942.

Wegner, D. M. (2004). Précis of the illusion of conscious will. *Behavioral and Brain Sciences*, *27*(5), 649-659.

Werkmäster, M. J. (2023). *Aspects of Blame: In which the nature of blame, blameworthiness, standing to blame and proportional blame are discussed* [Doctoral Thesis (monograph)]. Department of Philosophy, Lund University.

Wiegel, V., & van den Berg, J. (2009). Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, *1*, 233-242.

Wierenga, E. (1983). A defensible divine command theory. *Noûs*, *17*(3), 387-407.

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature*, *308*(5955), 181-184.

Wilson, E. O. (1971). *The insect societies*. Cambridge, MA: Harvard University Press.

Wilson, T. D. (2004). Strangers to ourselves. In *Strangers to Ourselves*. Harvard University Press.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2023). Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

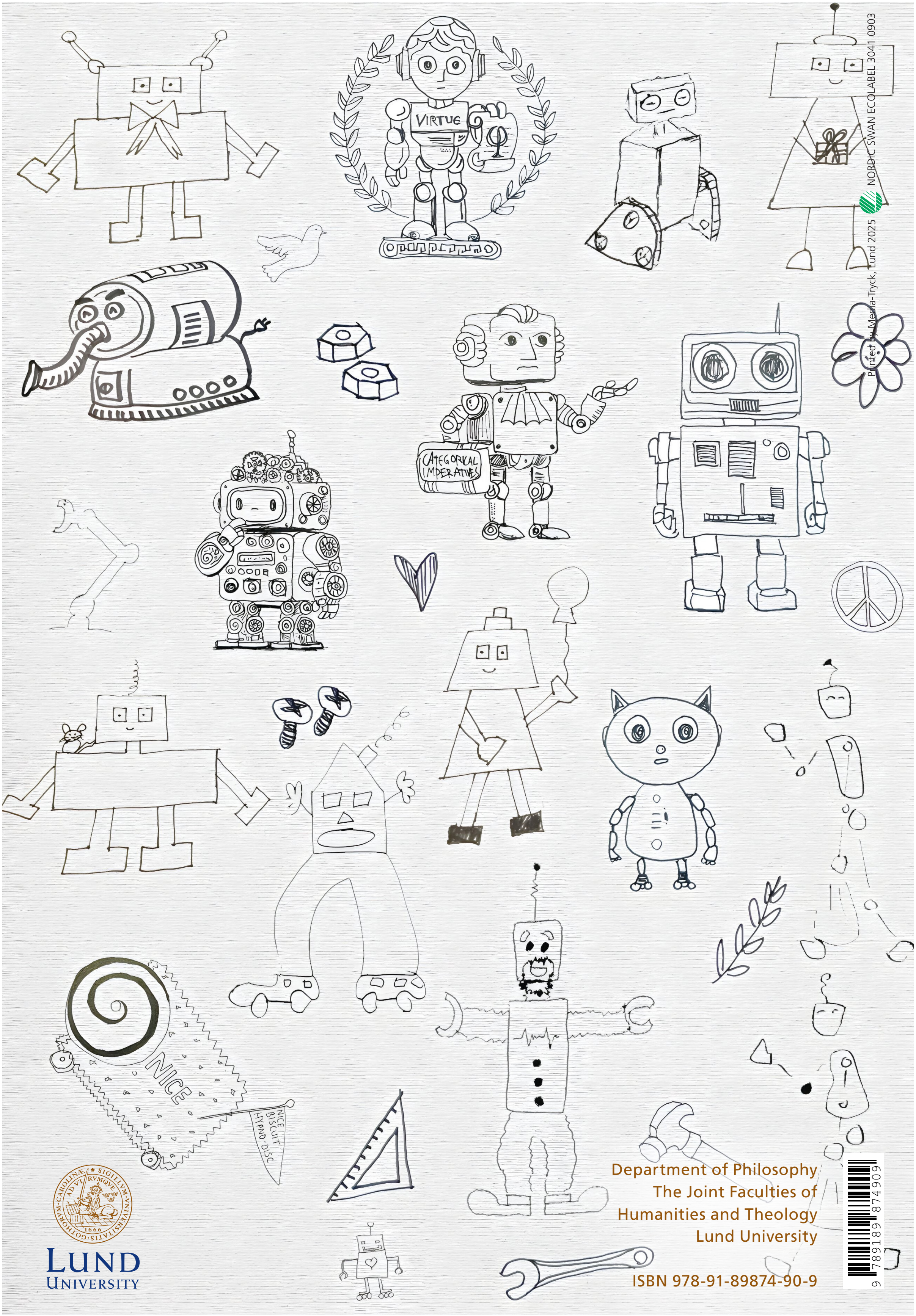
Wu, Y.-H., & Lin, S.-D. (2018). A low-cost ethics shaping approach for designing reinforcement learning agents. *Proceedings of the AAAI conference on artificial intelligence*, *32*(1).

Yampolskiy, R. V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. CRC Press.

Zagzebski, L. (2010). Exemplarist virtue theory. *Metaphilosophy*, *41*(1-2), 41-57.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Publications



Department of Philosophy
The Joint Faculties of
Humanities and Theology
Lund University

ISBN 978-91-89874-90-9

Lund University

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903