Artificial Moral Agency:

Autonomy and Evolution

Seyed Zacharus Gudmunsen

Submitted in accordance with the requirements for the degree of Doctor of Philosophy in Philosophy

University of Leeds

Inter-Disciplinary Ethics Applied Centre

School of Philosophy, Religion and History of Science

August, 2023

The candidate confirms that the work submitted is their own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Acknowledgments

This thesis benefited greatly from the support and guidance of my supervisors, Dr. Rob Lawlor and Dr. Graham Bex-Priestley. Rob has unfailingly offered detailed and valuable comments and criticism that have aided not just the arguments in this thesis, but my understanding and approach to philosophy more generally. Graham has always offered valuable insights into my arguments, showing me where things might be simpler or more complex. Both have been encouraging, understanding, and patient in their supervision, which I am grateful for.

I would like to thank the graduate students and faculty at the Inter-Disciplinary Ethics Applied (IDEA) Centre and the School of Philosophy at the University of Leeds. Both communities were warm, insightful, and generous. Also, those not at the University of Leeds, particularly communities in AI ethics and machine ethics, have been a valuable source of expertise and camaraderie. Especially the members of the Ethics and Technology Early-career Group (ETEG).

I also thank Dilara, for whom the writing of this thesis was just as punishing, but less rewarding. Without her, it would have been impossible, so I am glad she is here.

Finally, my family and friends, who have been equal parts support and distraction, thanks for your enjoyable company in this process.

Abstract

This thesis aims to establish the possibility of, and a pathway to, artificial moral agents. Artificial moral agents are argued to be of value not just for their practical performance, but because they offer a non-human perspective that can be used to make human theories more objective. The thesis works to a definition of moral agency, arguing that moral agents need to be intentional, morally reasons-responsive, and autonomous, but not necessarily conscious. Then, applying this to artificial agents, it draws on literature from moral epistemology and responsibility to argue that artificial agents normally fail to meet these criteria because they are not simultaneously morally reasons-responsive and autonomous. Following this, it argues that the most promising means of developing artificial moral agents is for artificial agents to evolve into moral agents. Even if not moral agents precisely, the evolutionary development suggested seems likely to produce autonomous artificial agents that respond to *some* moral reasons, which would still offer the desired non-human perspective.

Table of Contents

List of Illustrations

1. Introduction

Excerpt from 'Minds, Brains, and Programs' – John Searle, 1980

"Could a machine think?"

The answer is, obviously, yes. We are precisely such machines.

"Yes, but could an artifact, a man-made machine think?"

Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. It is, as I said, an empirical question.

"OK, but could a digital computer think?"

If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think

"But could something think, understand, and so on solely in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?"

This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.

"Why not?"

Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.

1.1. Looking for an Artificial Viewpoint

I am about to offer a bird's eye view of the thesis, presenting the detail of the thesis's arguments and organisation. Before that, I want to outline some very broad motivations. My motivations in writing on this topic are philosophical. I am not trying to solve problems (moral or otherwise) 'out there' in the world; nor do I expect to offer theoretical scaffolds for those faced with technological moral dilemmas. But I am writing about artificial systems[1] because I think they will be key to future philosophical thought. To be clear, I do not think that they will be key because humans will use them for genetic enhancement, anti-aging, full automation, or provoking catastrophes - although these possibilities are certainly interesting. I think artificial systems are key because humanity is crying out for an alternative viewpoint and artificial systems may yet be our best shot at providing one. I take the liberty of a few paragraphs here to explain, before getting on with the more rigorous business of the bird's eye view.

Take, as much as you can, the 'view from nowhere' (Nagel, 1989). See the world, as much as you can conceive of it, objectively. If you are a Kantian, imagine the noumena; if you are a scientist, imagine you understand the perfect scientific explanation of the universe; if you are a theist, imagine God's perspective; if you are an atheist, imagine the ideal observer's perspective. Of course, we humans are limited in our ability to do this. Even our strongest and best supported beliefs are coloured by our human nature and lived experience, no matter how we might try to inoculate ourselves against their influence.

Kant used transcendental arguments to reason from beliefs about experience to beliefs about experience's necessary conditions. Adopting the 'view from nowhere' requires pushing this all the way to the limit. For an agent who takes the 'view from nowhere', the necessary conditions for experience are on equal footing with the experience itself. All facts are flattened to a plateau – everything that is true is equally and obviously true and everything that are false is equally and obviously false. It is a philosopher's paradise. In fact, as far as I see it, all philosophy works to this end – to break down the mistaken and leave only the

[1] Throughout the thesis, I use 'artificial systems' to refer generally to artifacts which may or may not be agents. It might helpfully be understood to refer to 'artificial intelligences'. But I do not enjoy the connotations of 'intelligence' in that term, and 'artificial systems' is a wider concept (though that is rarely relevant).

glittering truth; to perfectly capture both simple and complex in one move. Call this destination 'objectivity'. I think almost every philosopher wants to get there.

How might we become more objective? The obstacle, it seems to me, is human nature. Our nature taints not only our individual beliefs, but our evaluative tools, our organisational structures, and our very methodologies. We are, by nature, subjective creatures, not objective ones. By now, the problem of *individual human subjectivity* is historical – we have established methods (science, philosophy, and art) and techniques (social organisation, institutions, and communication technology) for distilling individual human subjectivity into an intersubjective reality. (Evidently, this process is yet to be satisfactorily completed.) Using these methods, we aim to root out bias, mistakes, and prejudice. We aim see the world as a single entity illuminated by our collective epistemic effort.

The collective epistemic effort of human science and society yields a *more* objective view than the lone individual, but it is still far from true objectivity. The obstacle preventing us from moving further towards objectivity is still, I think, human nature. Science and society have been largely successful and offered great leaps of progress, but progress towards objectivity will (if it has not already) hit the roadblock of humanity's homogeneity. Trying to make the human species more objective by appealing to *more* humans is like trying to dry yourself with a wet towel. The human individual becomes more objective by collaborating with others. For humanity, seen as an individual species, the same principle applies. Humans need others to become more objective. Humans need *non-humans* to help triangulate, and ultimately correct our distinctively human biases, prejudices, and errors.

Extra-Terrestrials would be nice. Especially if they are radically different to humans. Xenobiology is, at least on this understanding of the nature of philosophy, one of the most important disciplines. But while we wait for aliens to arrive, we can turn to other things. Artificial systems are probably less useful than aliens for lighting the path to objectivity. But they are probably more useful than adding a further spoonful of humans to an already overpopulated epistemic stew.

To help humanity be more objective, artificial systems need to be distinct from them. That is, to offer the counterpoint humans need, artificial systems need their own nature. If artificial systems have their *own* nature, they can contrast and cross-reference their thinking and values with humanity's – promoting a more objective perspective that assumes neither the human nor the artificial worldview. Here, belatedly, we turn to artificial moral agents. Artificial

moral agents would have their own values, and thus may serve some small purpose in uncovering what is objective. Having your own *moral* nature seems an *essential* part of offering a non-human standpoint. In a narrower sense, being able to understand a non-human moral epistemology is likely to yield a metaethical bounty. Developing artificial moral agents is a worthy subgoal in developing the non-human standpoint that (I think) humanity needs, and outlining how to do so is the goal of this thesis.

1.2. The Bird's Eye View

This thesis concerns the possibility of artificial moral agents. It argues that artificial agents can be moral agents in principle, but that no artificial agents are yet moral agents. The reason for this is that moral agency requires agents to be both autonomous and (adequately) morally reasons-responsive. Traditional artificial systems cannot be both. I argue that the most promising way to develop artificial moral agents is to use evolutionary forces to do so.

The thesis attempts three interconnected tasks. First, identifying the conditions of moral agency and explaining why we ought to want advanced artificial systems. Second, assessing the extent to which most types of artificial systems can meet the conditions of moral agency. Third, outlining how to design artificial moral agents.

Part I concerns *why* we ought to pursue advanced artificial systems like artificial moral agents, and *what* moral agency is. In chapter 2 I offer a novel reason why we ought to pursue advanced artificial systems: they would increase cognitive diversity, benefitting human theories and understanding. Chapter 3 outlines my definition of moral agency: a moral agent is an agent that can adequately respond to moral reasons and that is generally responsible for their actions. Chapter 4 and 5 consider two properties which some think essential to moral agency: consciousness and intentionality. In chapter 4 I argue that consciousness is not a necessary condition of moral agency. In chapter 5 I argue that intentionality is necessary for moral agency and that artificial systems are intentional systems.

Part II of the thesis centrally concerns the ability of artificial systems to satisfy the conditions of moral agency. In chapter 6 I argue that artificial systems can, in principle, be morally reasons-responsive with an example. However, that artificial system, and many other types, are not responsible because they are reliant on moral deference. In chapter 7 I argue that

designed systems are generally unable to be responsible because they are not autonomous, but that some specific types of design may be able to create autonomous designed agents.

Part III of the thesis concerns evolution and its potential to produce artificial moral agents. In chapter 8 I outline how humans evolved to be morally reasons-responsive. In chapter 9 I sketch how an evolutionary artificial system, the likes of which are being developed in the field of 'artificial life', can evolve to be both autonomous (and thus responsible) and morally reasons-responsive.

1.3. Chapter Roadmap

Part I: What is a Moral Agent and Why Should We Want Artificial Ones?

Chapter 2. The Value of Advanced Artificial Systems and Cognitive Diversity

Chapter 2 concerns whether we ought to be motivated to generate advanced, independent artificial systems in the first place. I survey a series of reasons for and against the development of advanced artificial systems and develop a novel reason why developing them would be in the interest of humanity. That argument is that advanced artificial systems would increase 'cognitive diversity', which, if used appropriately, can lead humans to better, more robust, and more accurate theories and problem-solving.

I review several bodies of evidence that suggest that increases in cognitive diversity can be beneficial for human groups' theories and problem solving. Then I argue that advanced artificial systems would increase cognitive diversity in unique ways, thus offering unique benefits to human groups that include them. Whether this increased cognitive diversity is likely to be *used* by those groups in the best way to reap its rewards is unclear. But, in any case, one ought, all things considered, view increases in cognitive diversity and the potential improvements to theories and problem-solving as a significant pro tanto reason for developing advanced, independent artificial systems.

Chapter 3. Conditions for Moral Agency

In chapter 3, I outline the concept of agency that I use and set out the conditions of moral agency I will be using. I use a broadly functional account of agency such that an agent is an embodied system that exists in and interacts with an environment in a systematic way. I describe these conditions in more detail in the chapter itself, but one point to bear in mind is

that a system does not need to have intentional states to be an agent. Agents must, however, have some capacity for representation and motivation (functionally construed).

The central task of the chapter is to sketch a concept of moral agency. I suggest two necessary conditions for moral agency: to be a moral agent an agent must have the capacity to adequately respond to moral reasons and the capacity to be (generally) responsible.

These conditions, I argue, are largely uncontroversial. The first condition, the capacity to respond to moral reasons, is universally agreed upon. To add detail to the condition, I explain the sense in which I understand moral reasons, and what it takes to adequately respond to them. Some doubt the second condition. They think that moral agents need not be responsible. But I argue that moral agency as it is usually ascribed does include a responsibility condition, and it is this standard conception of moral agency that I target. I outline the standard definition of responsibility as requiring epistemic and control conditions. According to these conditions, an agent is responsible if it is *aware* of the relevant consequences of its actions and *in control* of the performance of the action.

Chapter 4. Consciousness Conditions for Moral Agency

Some think that moral agency has a 'Consciousness Condition' such that an agent must be conscious to be a moral agent. Chapter 4 argues against this claim. First, I outline the claim in the artificial moral agency literature that consciousness is necessary for moral reasons-responsiveness. Then I present the primary counterargument: the 'Epistemic Challenge'.

The Epistemic Challenge holds that consciousness cannot be a condition of moral agency because humans, despite not knowing whether other agents are conscious, competently ascribe moral agency. I discuss two counterarguments against the Epistemic Challenge: first, it is reasonable to think that humans do know that *other humans* are conscious when they ascribe moral agency; and second, concepts can have conditions that are unknowable or very difficult to verify.

I argue that both issues can be resolved by modifying the Epistemic Challenge to claim that humans do not have knowledge of whether *non-humans* are conscious. This modification, I suggest, can offer replies to both counterarguments while maintaining the conclusion that an agent can be morally reasons-responsive independent of being conscious. I offer further evidence for the Epistemic Challenge against non-humans with reference to arguments that non-human consciousness is indeterminate.

Chapter 5. Varieties of Functionalism for Artificial Moral Agency

In Chapter 5 I assess positions on whether moral agency requires *intentionality*. I discuss two varieties of 'functionalism' about moral agency: 'machine ethics functionalism' claims that an agent needs to meet functional or behavioural conditions to be a moral agent; what I call 'intentional state functionalism' claims that agents need to have the right intentional states to be moral agents, where their functional states just *are* intentional states.

I argue that intentional state functionalism is the most attractive functionalist account of moral agency. I defend this with a challenge for other functionalist accounts of moral agency: I consider the popular explanation that moral agents must have 'external' functional intentional states but not 'internal' intentional states. But, I argue, if there are such things as 'internal' intentional states then they are plausibly necessary for moral agency. Machine ethics functionalist accounts' best reply to this, I suggest, is to commit to intentional state functionalism and believe that there is only one type of intentional state, which is both functionally defined and necessary for moral agency.

I develop intentional state functionalism for artificial agents with an analogy to collective agency. In theories of collective agency is widely thought that intentional states are best understood as functional states which can be suitably interpreted as intentional states by others. Thus, an artificial agent has intentional states when it has the right functional states – and it seems that many artificial agents already do. They therefore satisfy any intentionality condition for moral agency. However, an important difference between collectives and artificial agents is that collectives are responsive to moral reasons because they are constituted by human members. Artificial systems do not have such resources to draw upon. I agree with Hess and Bjornson (2017) that to be morally reasons-responsive is a functional state. That does not, however, settle whether artificial systems can perform that function. In the following chapter I argue that they can, but that artificial agents struggle to be *both* morally reasons-responsive *and* capable of responsibility.

Part II: Troubles with Autonomy & Moral Reasons-Responsiveness

Chapter 6. The Moral Decision Machine and Artificial Moral Deference

Chapter 6 argues that some types of artificial systems can respond to moral reasons in a functionally equivalent way to humans. However, it then argues that they cannot do so while

being responsible. The chapter's central example is the 'Moral Decision Machine'. A machine that (discriminatively) applies a database of human moral decisions to new situations. The moral decision machine, I argue, responds to moral reasons in just the same way as humans do.

However, I argue that the Moral Decision Machine is reliant on moral deference. There is strong evidence from moral epistemology that reliance on deference for moral belief is problematic. The Moral Decision Machine, despite being sufficiently morally reasons-responsive, always defers. I generalise this to suggest that many artificial agents systematically defer.

One of the suggestions for why deference is problematic is that by deferring an agent undermines their *autonomy.* An agent that lacks autonomy cannot be responsible, and therefore cannot be a moral agent. So, deferential artificial systems cannot be moral agents even if they are morally reasons-responsive. I discuss whether *any* designed system can be autonomous in the following chapter.

Chapter 7. Can Responsible Agents be Designed?

In Chapter 7, I assess the Design Hypothesis: designed agents cannot be moral agents because they fail autonomy's *historical* condition. The chapter argues that the Design Hypothesis is false, but that many cases of design do undermine autonomy. First, I define 'design'. Then, I outline the historical account of responsibility, which claims that an agent must have a certain type of autonomy-promoting history to be responsible.

Many think that responsibility has a historical condition because of a series of cases known as 'manipulation cases'. To evaluate the Design Hypothesis, I consider 'design cases'. These are instances of design that appear to be autonomy-undermining like manipulation cases are. The standard position in machine ethics is that designed agents are just as responsible as humans. They think that designed agents cannot have their autonomy undermined because their histories are not meaningfully different to humans, who are autonomous.

I reply to this by defining the historical condition of autonomy. An agent has an autonomy-promoting history when the causal source of their actions is not another agent. To come to this definition, I combine Deery & Nahmias' (2017) causal interventionist and Waller (2014) and Liu's (2022) 'another-agent' approaches to historical conditions. With this definition, it is

apparent that designed agents are often not autonomous because the causal source of their actions is often their designer.

However, according to this condition, the Design Hypothesis is false because there are possible designed agents who can have autonomy promoting histories. Designed agents can have autonomy-promoting histories if their actions' causal sources are non-agent forces. Two possibilities of non-agent causal sources are randomness and evolution – I suggest that evolutionary forces are the most promising means of designing artificial moral agents.

Part III: The Evolutionary Pathway to Artificial Moral Agency

Chapter 8. The Evolution of Moral Capacities

In chapter 8, I describe the evolution of human moral reasons-responsiveness by drawing on theories developed by Philip Kitcher (2011), Michael Tomasello (2016), and Richard Joyce (2006), among others. As outlined in chapter 3, moral reasons are seen as reasons that can plausibly be used in a moral theory. Using moral reasons in this way entails that some simple organisms are morally reasons-responsive. However, moral agency requires *adequate* moral reasons-responsiveness – bees (for example) are morally reasons-responsive but not *adequately* moral reasons-responsive.

The chapter divides capacities for moral reasons-responsiveness into three types: biological, psychological, and conceptual. Humans use all three. Biological capacities are often seen as a foundational capacity in the evolution of moral reasons-responsiveness but lack sufficient nuance to be *adequately* moral reasons-responsive. Psychological capacities are likewise an important evolutionary stage in the development of adequate moral reasons-responsiveness, but humans are only adequately morally reasons-responsive because they have conceptual capacities too.

Biological and psychological capacities are important developmental stages for adequate moral reasons-responsiveness. But they are also relatively common in nature. The bottleneck in nature is conceptual capacities. I outline some theories about the environmental and evolutionary conditions that led to protohumans developing conceptual capacities for moral reasons-responsiveness.

I suggest the project of evolving an artificial agent with adequate moral reasons-responsiveness ought to aim to evolve agents with functional equivalents of all three biological, psychological, and conceptual capacities. When these capacities are combined, as they are in humans, they are proven to result in adequate reasons-responsiveness.

Chapter 9. Artificial Life and Moral Agency

Chapter 9 argues that artificial life, that is, the evolutionary development of artificial systems, can generate artificial agents which are both morally reasons-responsive and responsible. I explain several challenges for designing autonomous artificial life agents. Chief among them the challenge of being sufficiently 'hands off' to design artificial agents with autonomy-promoting histories. While difficult, I suggest that this challenge can be overcome.

Designing artificial life organisms that evolve adequate moral reasons-responsiveness is more challenging. Biological capacities can be readily evolved by artificial life systems. But psychological capacities are more challenging. I outline some options for overcoming these challenges and evolving artificial life with psychological capacities. But it is even harder to evolve *conceptual* capacities. There does not seem to be a readily available and robustly effective means of developing these capacities through evolution (on Earth only one species has achieved this so far). I suggest some non-evolutionary solutions to this, acknowledging that they may interfere with autonomy.

I conclude that artificial life can evolve to be artificial moral agents. Though there are clearly many practical challenges in achieving this. I discuss how some partial successes could be useful and valuable, should artificial moral agency fail to evolve. Crucially, evolved artificial agents with (perhaps inadequate or insufficiently autonomous) moral reasons-responsiveness should provide the kind of non-human standpoint useful for aiding humans develop more objective theories.

Part I: What is a Moral Agent and Why Want Artificial Ones?

2. Cognitive Diversity and Advanced Artificial Systems

2.1. Introduction

It is unclear whether humans can develop advanced[2] artificial systems. It is perhaps equally unclear whether humans *should* develop them. Until a few years ago, most philosophers in the field of 'machine ethics'[3] assumed that we should. It was thought that advanced artificial systems would lead to better understanding of the world and better material consequences for its inhabitants, especially if its designers were cautious enough to program in some ethical guidelines. However, recently, some have claimed that humans should not develop advanced artificial systems even if they can. They defend this claim in two ways: first, by arguing that the positive reasons for developing advanced artificial systems are unconvincing, second, by arguing that the positive reasons for developing advanced artificial systems are outweighed by negative ones.

I do not intend to settle the question here. Deciding whether humans should try to develop advanced artificial systems is complex and depends on many theoretical and practical reasons, the unveiling of which require significant concerted research. I here contribute to that by presenting a novel reason for wanting advanced artificial systems. The argument establishing this reason is that advanced artificial systems increase 'cognitive diversity', that increased cognitive diversity is beneficial for humans.

The structure of the chapter is the following. First, I describe the context of the debate about whether we should develop advanced artificial systems in terms of the reasons for and against. Then, I develop the 'cognitive diversity' reason for developing advanced artificial systems. The argument begins with several pieces of evidence for the idea that cognitively diverse groups can produce better performing, more robust, and more accurate theories than their non-diverse counterparts.

---

[2] 'Advanced' meaning with human level or above performance in many areas.
[3] The field that studies the possibility of and means of development for artificial moral behaviour.

I present three bodies of evidence for the claim that cognitive diversity results in better decision making in groups. First, cognitive diversity is epistemically beneficial because it provides a greater variation of evidential sources, which provide more robust and accurate theories. The chief example of this is in the epistemic benefits of social diversity, which is argued for in the discipline of feminist epistemology. Second, agent simulations in philosophy of science offer evidence that cognitively diverse groups produce better theories and solve problems more effectively than homogenous groups. Third, empirical evidence from attempts to promote cognitive diversity (and/or social diversity) offer a tentatively positive effect on the outcomes of diverse groups' decisions.

Having offered evidence for cognitive diversity being beneficial for groups, I then argue that advanced artificial systems will increase cognitive diversity in a group they are part of. Along the way, I compare the expected cognitive diversity of human groups, human/non-advanced artificial systems hybrid groups and human/advanced artificial system hybrid groups. I claim that that human/advanced artificial system hybrid groups have distinctly greater cognitive diversity compared to the others. If this is correct, then one good reason for developing advanced artificial systems is that they will promote cognitive diversity. Of course, this is only a single pro-tanto reason, and may not be sufficient to tip the balance of the overall decision to make advanced artificial systems but should nonetheless be an important part of that decision.

2.2. Does Anyone Know If We Should Make Advanced Artificial Systems?

In the philosophical literature there are two primary sources of reasons against wanting advanced artificial systems. First, theoretical challenges to the project of 'machine ethics' – that is, the project which aims at the design and implementation of artificial moral agents – and second, the 'alignment problem', a problem most often discussed in the context of the existential risk of super-intelligent artificial systems. In response, those who support the project of machine ethics provide reasons for wanting artificial moral agents, and others have argued that the alignment problem should not deter us from developing advanced artificial systems.

Artificial moral agents are not necessarily advanced artificial systems. However, the meeting the conditions for being a moral agent are very likely to require an agent to be 'advanced', so reasons against wanting artificial moral agents are likely to also reasons not to make advanced artificial systems.

The discussion here will overview the reasons involved without assessing their relative effectiveness, as I am concerned here with presenting a new reason rather than evaluating the reasons already on the table. However, the reasons about to be presented seem not to definitively settle the question of whether we should want advanced artificial systems, and therefore the exploration of new reasons seems a worthwhile exercise.

What follows are several reasons for *not* wanting advanced artificial systems. These are presented by those who believe that developing advanced artificial systems would result in harm or would be unethical in and of itself and that therefore we should not continue to develop advanced artificial systems (at least until the problems they discuss are resolved). Call those who take this view 'advanced artificial systems pessimists'. Call those who disagree and think that we *should* continue to develop advanced artificial systems 'advanced artificial systems optimists'.

From the machine ethics side of things, Ryan Tonkens' 'A Challenge for Machine Ethics' (2012) argues that it would be unethical, according to Kantian moral theory, to produce a Kantian artificial moral agent. The reason is that in designing artificial moral agents, Tonkens says, we are inevitably going to treat them as *means* rather than *ends.* Therefore, for a Kantian to design an artificial moral agent must commit a wrong to do so. From a Kantian perspective then, we should not develop artificial moral agents because doing so would be unethical.

Joanna Bryson (2010, 2018) offers a more practical reason to refrain from developing advanced artificial systems – advanced artificial systems would be likely to be moral patients, and the existence of artificial moral patients imposes moral duties on humans that are likely to be hard to recognise and fulfil.

A popular topic in this area is 'responsibility gaps'[4] (Matthias, 2004; Sparrow, 2007). There is a 'responsibility gap' when an artificial agent commits a harm without any agent (either artificial or human) being responsible for it. Developing advanced artificial systems that are incapable of being responsible for their actions would increase the possibility (and probably severity) of responsibility gaps occurring, and therefore potentially result in harmful actions which, problematically, no-one can be held responsible for. Responsibility gaps are a reason against wanting advanced artificial systems because with if no one is responsible for a harm, then moral and legal practice cannot minimise it.

Two more general reasons for pessimism are that, first, advanced artificial systems would 'supercharge' the harm caused by harmful decisions. Suppose a designer or artificial system makes a mistake or forms an evil intention. The potential harms may be dramatically increased through having a powerful artificial system act based on this mistake or evil intention. Developing advanced artificial systems may be compared to the development of stakes-raising technology like nuclear missiles. Increasing the stakes of our moral actions is unlikely to be desirable, even if there may be expected benefits. Second, advanced artificial systems may lead to greater harms through humans misunderstanding their abilities, or by changing society. Possible societal changes such as technocracies or fully automated worlds of work may result in reductions of welfare or systems of oppression. Once again, while advanced artificial systems may facilitate great good by changing society, they may also facilitate great harm; and we have good reason to be cautious about raising the stakes.

Finally, one area of study focuses on the 'existential risk' that advanced artificial systems may lead to (Bostrom, 2016; Gabriel, 2020; Ord, 2020; Turchin & Denkenberger, 2020; Vallor, 2016). One line of thought suggests that advanced artificial systems may be able to improve themselves exponentially, leading to an 'intelligence explosion' where they rapidly develop into superintelligences. This would raise the stakes even further, since superintelligent agents could cause *even more* harm or good, and could not be easily controlled. One program of research that attempts to mitigate these problems is in 'AI Safety' or 'value alignment'. These researchers, which include machine ethicists and computer scientists, aim to generate design principles for the development of advanced artificial

[4] And related 'retribution gaps' (See Nyholm, 2017).

systems that won't cause harm. They take the risks of advanced artificial systems very seriously and consider value alignment to be an ongoing and immature research project. Typically suggesting that development of more advanced artificial systems should be halted until better design principles can be created.

There are several reasons to *not* want advanced artificial systems. I now turn to advanced artificial systems optimists, who disagree with many of these reasons. As mentioned, I will not consider the *balance* of the reasons here but note that the significant disagreement involved suggests that the matter is far from settled.

Advanced artificial systems optimists present their own positive reasons for wanting to develop advanced artificial systems. Aimee van Wynsberghe & Scott Robbins argue that none of the reasons machine ethicists presented for developing artificial moral agents are convincing (2019). Formosa & Ryan (2021) argue against this.

The best reasons for developing advanced artificial systems are obvious. It is evident, for example, that advanced artificial systems can aid in human efforts. Their ability to process information can automate both labour intensive and cognitively complex processes, offer evidence for theories, and otherwise aid human reasoning. This is, prima facie, a reason for wanting them.

One notable question is whether these advantages are limited to advanced artificial systems or whether they can apply just as well to less advanced artificial systems. If there is no gain from developing advanced artificial systems over non-advanced artificial systems, then there is no corresponding reason to develop advanced artificial systems. One response is that artificial systems already benefit humanity, and that artificial systems will increase that benefit. In any case, the reason for developing advanced artificial systems that I present *is* unique to advanced artificial systems.

It has been suggested that artificial moral agents are needed to prevent harm that powerful but non-advanced artificial systems would otherwise cause (Allen et al., 2006; Moor, 2006; Scheutz, 2016; Sison & Redín, 2023; Wallach & Allen, 2008). Since it is inevitable that artificial systems will become increasingly powerful, increasingly powerful means of controlling them are necessary, and one way of doing this is for them to be morally competent themselves. So, one reason to want advanced artificial systems is to keep less advanced artificial systems in check.

The case has also been made that artificial moral agents will allow us to learn more about ethics and contribute towards ethical theories (S. L. Anderson & Anderson, 2020; Ioan & Howard, 2017; Wiegel, 2006). I will return to this idea in the conclusion, as the reason for developing advanced artificial systems presented here offers further evidence for thinking that advanced artificial systems would improve our ethical theories. If so, I think this a convincing reason to want them.

Owe, Baum and Coeckelbergh (2022) offer some possible reasons to think that advanced artificial systems would be intrinsically valuable. They suggest that certain types of advanced artificial systems may be intrinsically valuable in being subjects of experience, promoters of diversity, aesthetics, or being alive. Though these do not initially seem to offer a reason for wanting advanced artificial systems, if we ought to be motivated to create more beings with intrinsic value (pace Bryson), then they may be. Owe et al. discuss promoting *diversity* as a possible intrinsic value; an idea that I am sympathetic to and generate a version of in this chapter. Though they focus on *biodiversity* while I focus on *cognitive diversity*.

Finally, some of those who work in value alignment and AI safety do so because they anticipate that alignment problems will be solved or do not need to be solved. Assuming an intelligence explosion is possible, some suggest a superintelligent artificial system would be super-moral (Dietrich, 2001). While this is little more than speculation, if an intelligence explosion would necessarily lead to better moral outcomes and advanced artificial systems would increase the chance of an intelligence explosion, then we may have a substantial reason to develop advanced artificial systems.

So, advanced artificial systems optimists can also offer a variety of reasons for thinking that we ought to develop advanced artificial systems. Though none of the reasons on either side seem to be conclusive. The question of whether we ought to create advanced artificial systems depends on the balance of reasons. The introduction of new possible reasons, therefore, can better inform us about whether we ought to be pessimists or optimists. As mentioned, I do not intend to *resolve* the decision here, but instead to suggest that there is a reason for advanced artificial systems optimism that has been overlooked.

The cognitive diversity reason: Advanced artificial systems offer an epistemic advantage to human-only groups by increasing their cognitive diversity, therefore, humanity has a reason to create advanced artificial systems.

The cognitive diversity reason's effectiveness depends on two premises:

(1) Greater cognitive diversity offers an epistemic advantage to a group.
(2) Advanced artificial systems increase cognitive diversity in a group.

I will now argue for these premises. The next three sections (2.3.-2.5.) explore and defend (1), and 2.6 argues for (2).

The following sections argue that (1) is true by discussing the advantages of cognitive diversity in various contexts. This discussion not only provides evidence for (1) but helps to narrow down which types of cognitive diversity are beneficial and when they are most useful.

2.3. The Value of Cognitive Diversity

The overall idea here is that groups which are cognitively diverse, such as a group of a man, a woman, and an alien; or a group of a scientist, a doctor and an artist, have a demonstrable and often desirable epistemic advantage over a group which is cognitively homogenous, such as a group of three men with the same socio-cultural background. Given this, I will argue, a group of two humans and an advanced artificial system has a similar desirable epistemic advantage over a group of three humans. This results in the cognitive diversity reason being a good one. The first stage of the argument is to argue for (1) – that cognitive diversity offers an epistemic advantage.

Several of the reasons for thinking that cognitive diversity is beneficial that I present are informed by Scott Page's work on diversity (2007). Page defines cognitive diversity as variation within a group's *perspectives, interpretations, heuristics,* and *predictive models*.

Page summarises them as:

"Diverse Perspectives: ways of representing situations and problems

Diverse Interpretations: ways of categorizing or partitioning perspectives

Diverse Heuristics: ways of generating solutions to problems

Diverse Predictive Models: ways of inferring cause and effect" (Page, 2007, p.7)

Essentially, cognitive diversity is represented by variation within a group's *beliefs, aims, access to evidence, problem-solving methodology* and *conceptual frameworks*.

This section argues that cognitive diversity can provide an epistemic advantage for a group. By 'epistemic advantage', I mean an increase in a group's access to evidence and/or an increased ability to process evidence within a decision-making process. Page (2007, p. 13) puts this as that "[d]iverse perspectives and tools enable collections of people to find more and better solutions and contribute to overall productivity".

Before heading into the evidence for the claim, I, again in the footsteps of Page (2007, pp. 13-4), clarify the relationship between cognitive diversity and the related (and probably more familiar) notion of social diversity. Cognitive diversity positively correlates with demographic, gender, cultural, and other forms of social diversity. After all, differences in social backgrounds, cultures, beliefs, or experiences can (and often do) promote cognitive diversity. Arguments that social diversity offers an epistemic advantage can therefore be generalised to argue for the benefits of cognitive diversity (and vice versa).

But cognitive diversity is not simply a product of social diversity. For one, there may be social diversity without cognitive diversity: a socially diverse group may share the same beliefs, methodologies, concepts, and so on. Furthermore, cognitive diversity can be promoted by things other than social diversity: a socially homogenous group may be cognitively diverse if it includes members who have a variety of methodologies, conceptual frameworks, and beliefs. For example, a socially homogenous group that contains members with beliefs that range across the political spectrum has more cognitive diversity than a socially homogenous group that share political beliefs.

Cognitive diversity can be hard to pin down. There is no easy quantification that allows comparison of two similarly cognitively diverse groups. Cognitive diversity as understood here is broad and can be increased by numerous types of differences between group members. This is intentional. This paper's argument does not need an exhaustive and exclusive definition of what exactly constitutes 'cognitive diversity', so long as you have a general idea of some clear cases in which one group has greater cognitive diversity than another.

That said, one possible aid in measuring cognitive diversity is statistics. Particularly the range and variance of the differences between group members. A group with a greater range and/or variance of cognitive diversity variables is more cognitively diverse than one with a smaller range/variance. See also general definitions for diversity (Stirling, 2007) which invoke

similar statistical concepts to measure diversity that may be usefully applied to cognitive diversity[5].

So, what about the claim that greater cognitive diversity offers a group an epistemic advantage? I review three ways to defend the claim that cognitive diversity offers an epistemic advantage. First, Page and others defend the claim with models of cognitive diversity. Second, some within feminist epistemology defend the claim by arguing that diversity allows groups to be more objective. Third, the outcomes of diversity promoting initiatives (such as 'affirmative action') offers mixed empirical evidence for the claim that cognitively diverse groups are better decision makers. I will take these in turn.

One important qualification is that *offering* an epistemic advantage is not the same as *taking* an epistemic advantage. Even if cognitive diversity offers an epistemic advantage to a group, the group may not take the advantage. If cognitive diversity offers an advantage that can, in principle, be taken by a group without incurring disadvantages, then the claim that cognitive diversity offers an epistemic advantage is true.

Hong and Page (2004) investigated the relationship between cognitive diversity and better problem solving in groups. That they are related is not guaranteed, as cognitive diversity is distinct from problem solving ability. An individual may increase the cognitive diversity of a group while being a relatively poor problem solver; and another individual may be a very good problem solver but not contribute much to a group's cognitive diversity. Hong and Page's claim is that groups with high cognitive diversity on average perform better than groups with high problem-solving ability. Their tagline is 'diversity trumps ability'.

Hong and Page use agent-based models in which two groups of agents are compared. One group has high cognitive diversity and the other has low cognitive diversity. The experiments were repeated with various average group ability levels. Hong and Page found that even when a cognitively diverse group had lower ability levels, they often made better decisions. This result was backed up by a theorem, as Page (2007, p. 162) says: "it is a logical truth".

There are some conditions: first, the problems must be hard enough that there is no single solution many will come across – i.e., there is no benefit from cognitive diversity in easy

[5] Concepts used to measure biodiversity might be useful as a comparison too (Sarkar, 2010)

problems like 2+2. Second, all individuals must be relatively good at solving the problem at hand, as Page (2007, 160) says: "a collection of random people would not outperform a collection of top statisticians on a statistical problem." The individuals in the group must have a decent ability for solving the problem to be able to collectively overcome more able but less cognitively diverse groups. The other conditions are that there must be an individual that is able to improve on suboptimal solutions[6]; and that there is a relatively large pool of problem solvers such that there can be a meaningful difference between a cognitively diverse group and a highly able group.

Enabling conditions noted, the result is that "a randomly selected collection of problem solvers outperforms a collection of the best individual problem solvers." (Page, 2007, 162). This result led to quite a bit of further research (Some philosophically oriented examples are Grim et al., 2019; Holman et al., 2018; Peters et al., 2020; Steel et al., 2021). One point that arose is that for a cognitively diverse group to perform better, the group must be able to effectively include the diverse individuals in a decision-making process. In practice, as we will see, this is a difficult thing for a group to do. Page suggests that the best group is diverse, but not *so* diverse that communication and co-ordination becomes too challenging.

There are several indirectly related discussions about cognitive diversity, or other related forms such as 'varied evidence', in philosophy of science that generally support the claim that it offers an epistemic advantage (Landes, 2020; Muldoon, 2013; Osimani & Landes, 2023).

The argument just reviewed involves formal approaches, and multiple further explanations and conditions that I have not covered here. Nonetheless, I suggest that a reasonable takeaway is that cognitive diversity offers an epistemic advantage to a group. Of course, there are caveats: relatively specific conditions must be met. But especially if a group can incorporate cognitively diverse individuals into their decision making, there is good evidence that increasing a group's cognitive diversity can increase their problem-solving ability (and, Page argues, potentially their predictive power too (Page, 2007, Ch. 7-8)).

[6] This is less restrictive than it sounds, since in the real world this restriction is that there must exist an individual who can improve on the solution.

> [O]rganizations, firms, and universities that solve problems should seek out people with diverse experiences, training, and identities that translate into diverse perspectives and heuristics. Specifically, hiring students who had high grade point averages from the top-ranked school may be a less effective strategy than hiring good students from a diverse set of schools with a diverse set of backgrounds, majors, and electives. (Page, 2007, p.173)

Some cast doubt on the Hong & Page result. Some are skeptical from a mathematical point of view, and some believe that the conditions are restrictive enough to render the result of the epistemic advantage trivial and irrelevant to real world deliberative practice. While there have been responses to these criticisms, the mathematical nature of the model does not offer a especially convincing *explanation* of why cognitive diversity helps. Fortunately for defenders of the cognitive diversity's value, there is ample explanations of this sort in theories of Feminist Epistemology.

2.4. The Value of Cognitive Diversity in Feminist Epistemology

The preceding discussion has already offered evidence that cognitive diversity offers an epistemic advantage. To add to that, there is a large body of work in feminist epistemology that supports similar conclusions. Their claim is that diversity (understood broadly) promotes the examination of assumptions and thus results in more objective scientific theories.

Feminist 'standpoint theory' claims that various members of society have different 'epistemic standpoints'. An epistemic standpoint is something like what Page (2007) calls a person's 'toolbox' – that is, the combination of a person's perspective, heuristics, interpretations, and models. Within feminist standpoint theory the most valuable epistemic standpoints are often taken to be those that are *different*. Often, this difference stems from being part of a marginalised group:

This justificatory approach originates in Hegel's insight into the relationship between the master and the slave and the development of Hegel's perceptions into the "proletarian standpoint" by Marx, Engels, and Georg Lukacs. The assertion is that human activity, or "material life," not only structures but sets limits on human understanding: what we do shapes and constrains what we can know. As Hartsock argues, if human activity is structured in fundamentally opposing ways for two different groups (such as men and women), "one can expect that the vision of each will represent an inversion of the other, and in systems of

domination *the vision available to the rulers will be both partial and perverse.*" (Harding, 2016, p. 120, my emphasis).

A group containing differing epistemic standpoints, it is argued, contains a route to knowledge ('vision') that is absent (or ignored) in mainstream homogenous groups (Harding, 2016; Intemann, 2009; Longino, 2018; Wylie, 2012). Some agents' (particularly those who from marginalised groups) epistemic standpoints can render unseen assumptions 'visible'. Under standpoint theory, diversity, in terms of containing different epistemic standpoints, therefore offers a group an epistemic advantage in the practice of developing theories.

One suggestion for *how* this happens is that 'insider-outsiders': those who can understand the issues but are not part of the dominant narrative, can see things that others cannot:

> Her unique position as an ''insider-outsider'' provides her with both expertise and experience to recognize problematic background assumptions and to identify the sort of evidence that will be relevant given the aims of the research. Because she is an ''insider'' she has the relevant expertise to be able to understand and identify assumptions that are being made in her field. Yet as an ''outsider,'' or as a member of a group that has been historically excluded from such research, she has had experiences that allow her to identify the limitations and problems with some of those assumptions. In this way, scientific communities that include members of oppressed groups with experiences relevant to the research *can access a wider range of empirical evidence, more easily identify problematic background assumptions, and more readily generate new hypotheses, models, and explanations.* (Intemann, 2010, pp. 788–789, my emphasis)

Standpoint theory typically focuses on *social diversity* more than *cognitive diversity.* The claim tends to be that those with different social backgrounds have epistemic standpoints which offer insights in developing scientific theories. That said, "[c]ontemporary standpoint theorists, however, have denied that standpoints are merely socially located perspectives" and claim that a standpoint need not involve "a universally shared perspective of all members of a particular social group. Individuals may contribute to the achievement of a critical consciousness within an epistemic community in different ways." (Intemann, 2010, p. 785). So, epistemic standpoints can be distinguished from social identity. For this paper, the claim under scrutiny concerns cognitive diversity. It is that epistemic standpoints which *increase*

cognitive diversity are (also) valuable[7]. This, I suggest, benefits from the same arguments that standpoint theorists use to defend the value of marginalised groups' epistemic standpoints – since, as discussed earlier, social diversity (and the accompanying epistemic standpoints) correlate with cognitive diversity.

Additionally, the claim that cognitive diversity offers an epistemic advantage is not that far away from the position of some in feminist epistemology who refer to the desirability of 'epistemic diversity' (E. Anderson, 2006; Solomon, 2001, 2006). Feminist empiricists can be interpreted as supporting something along these lines. Intemann explains the Feminist empiricist position:

> [F]eminist empiricists have advocated for scientific communities comprised of individuals with diverse values and interests. Consensus that emerges from a community with diverse values will be more likely to be rational, rather than implicitly based on widely shared, erroneous moral or political values (provided the community also meets the other conditions for objectivity). In this way, inquirers with diverse values and interests provide a system of checks and balances so as to ensure that the idiosyncratic values or interests of scientists do not inappropriately influence scientific reasoning. (Intemann, 2010)

Feminist standpoint theorists and feminist empiricists disagree on several points – one of them being that standpoint theorists emphasise social situatedness while feminist empiricists emphasise diverse values. As such, while both positions are aligned with the claim that cognitive diversity offers an epistemic advantage, feminist empiricism is probably closer aligned.

For example, one explanation for why a group with 'diverse values and interests' can provide checks and balances because it is *cognitively diverse.* Cognitive diversity seems to be part of having broadly 'diverse values and interests' and reasonably offers 'checks and balances' in scientific practice for the same reasons that diverse values and interests do. That is, cognitively diverse groups can draw on a larger base of evidence, perspectives, experiences, and approaches that enable them to identify and avoid widely shared erroneous assumptions.

---

[7] To be clear, I am not arguing *against* feminist standpoint theory here, as mentioned earlier, social diversity correlates with cognitive diversity; and standpoint theory often targets standpoints in a nuanced way that is likely to adequately encapsulate cognitive diversity or its advantageous qualities. At the risk of oversimplification, I am suggesting that their arguments offer evidence for the cognitive diversity claim.

I leave the discussion there, though this is again only the briefest exploration of the issues involved. However, the conclusion that, in line with feminist epistemology's arguments, increasing cognitive diversity in a group offers an epistemic advantage by unveiling hidden assumptions and offering alternative viewpoints is appealing.

As a final note on the philosophical theories that support the claim. I want to point out that the idea that cognitive diversity offers an epistemic advantage ought to be appealing for traditional epistemologists, too. The basic idea is that more sources of evidence increase accuracy. Much of epistemology depends on something like this as a general truth. The general strategy of coherentist accounts (and one accepted by all but the most radical foundationalist accounts) of epistemology is that more evidence is better – and that more *types* of evidence are better than more of the same evidence (BonJour, 1988; Olsson, 2021). Since cognitive diversity allows a group to access more types of evidence, via the actions of cognitively diverse members who can present evidence which are theory-laden in a different way, then they have greater warrant for their beliefs. Thus, they are more *objective* – meaning their theories are, epistemically speaking, able to be more *accurate*. W. V. O. Quine (1960) is a key origin point for this kind of approach: the underdetermination of theory and the theory-ladenness of evidence pushes scientists and epistemologists to search for more evidence to increase their epistemic warrant. Cognitive diversity offers a route to that evidence, and thus ought to be desired by scientists and epistemologists. Of course, this is merely a crude sketch of the position – but it describes some of the epistemic foundations that feminist epistemology was built upon, and hints that those foundations are sympathetic to the cognitive diversity claim.

2.5. Empirical Evidence for the Value of Cognitive Diversity

There is longstanding debate about whether diversity offers advantages in practice – the so called 'diversity hypothesis'. In assessing this hypothesis, some investigate social diversity, while others more narrowly focus on cognitive diversity. Sujin K. Horwitz and Irwin B. Horwitz take a positive position on the claim and refer to a general definition of cognitive diversity:

> Cognitive diversity in the team context is defined as the degree to which team members differ in terms of expertise, experiences, and perspectives. Using the theoretical arguments of the cognitive diversity hypothesis, several researchers have

> argued that team diversity has a positive impact on performance because of unique cognitive attributes that members bring to the team. Ultimately, cognitive diversity among heterogeneous members promotes creativity, innovation, and problem solving, and thus results in superior performance relative to cognitively homogeneous teams. (Horwitz & Horwitz, 2007)

But not everyone is universally positive. Diversity is generally seen as a double-edged sword (Mello & Rentsch, 2015; van Dijk et al., 2012) that is not reliably beneficial. First, there are inconsistent results. Diversity appears to be advantageous in some contexts (For example, Beckman et al., 2012; Hoever et al., 2012; Olson et al., 2007). But cognitive diversity is arguably disadvantageous in other contexts or more generally. Katherine Williams and Charles O'Reilly (1998) reviewed eighty studies on demographic diversity and group performance and found that the effects of cognitive diversity were negative. Several others found that cognitive diversity had a negative or mixed relationship with group performance (Aggarwal & Woolley, 2018; Mello & Delise, 2015; Miller et al., 1998; Pieterse et al., 2011; van Dijk et al., 2012; van Knippenberg & Schippers, 2007).

Such results may seem to offer evidence against the claim that cognitive diversity offers an epistemic advantage. However, several compensatory factors should be noted. First, if, as Page claims, and several others have noted, increased cognitive diversity increases conflict within a group, then the expected result would be that the epistemic advantage of cognitive diversity was counterbalanced by increased conflict, which may lower performance. Clearly, if a group cannot work well together as a team, or agree on a strategy or methodology, then they will be less effective, even if their cognitive diversity may offer epistemic advantages in principle. Increased group conflict can also explain negative self-appraisals. This was borne out in one study, where diverse groups rated their performance negatively when it was actually just as good as in homogenous groups (Tyran & Gibson, 2008).

However, even group conflict is not necessarily negative for a group's performance. In another study, higher group conflict accompanied higher performance (Olson et al., 2007). Cognitive diversity has been reliably shown to improve a team's creativity and innovation (Hoever et al., 2012; Milliken & Martins, 1996), and one possible explanation for this is that creativity benefitted from cognitive diversity without being negatively impacted by group conflict.

So, if one believes the group conflict explanation for the mixed empirical evidence, then the cognitive diversity claim can still be true. The problem, it seems, is that group's struggle to

take the epistemic advantage, because cognitive diversity correlates with higher group conflict. Thus, while the mixed empirical evidence is disappointing for the cognitive diversity claim, it is not necessarily evidence *against* that claim.

Overall, the empirical evidence neither confirms nor denies the claim that cognitive diversity offers an epistemic advantage. The best evidence for it is theoretical support from philosophy of science and epistemology. This is not merely theoretic, either, there are several proposals for improving scientific practice that make good cases that diversity can benefit group decision making and outcomes (Cruz & Smedt, 2013; Intemann, 2015; Wylie, 2015).

In the following section I will assume that (1) is true, and argue for (2), that artificial systems can increase cognitive diversity, and that in particular advanced artificial systems can increase cognitive diversity in a unique way. Thus, I will argue that the cognitive diversity reason is a valid reason for developing advanced artificial systems.

2.6. How Artificial Systems Increase Cognitive Diversity

I here outline how artificial systems in general can increase cognitive diversity in a group that uses them. Later, I'll argue that advanced artificial systems increase cognitive diversity in a comparable but distinct way. Let us call human agents that are enhanced by artificial systems 'cyborgs'. In this sense, we are all cyborgs – I am enhanced in my epistemic capacity to access philosophical papers through using the artificial system of my laptop (and the larger artificial system of the internet). Cyborgs do not need to be high-tech, however, a person wearing glasses is also a cyborg. A cyborg is a human-artificial hybrid agent.

To use the terminology from feminist epistemology, being a cyborg seems to offer a different 'epistemic standpoint'. Imagine a team is competing in a quiz. Regular cognitive diversity can help them in the quiz: the various beliefs, strategies, and values of cognitively diverse members are more likely to lead to non-overlapping knowledge. Suppose, then, some of the team members are cyborgs who can access the internet. This is normally cheating, but only because of *how obviously* it offers an epistemic advantage.

So, here's the question: can the internet-cheating epistemic advantage be attributed to an increase in cognitive diversity? The path to a positive answer lies in the nature of the advantage. The internet-connected cyborg offers an advantage because they can access different sources of evidence, different methodologies, and different conceptual frameworks

compared to the other members of their team. Thus, the cyborg *does* increase cognitive diversity, and it is, it seems, the very same cognitive diversity promoting features that enable them to improve the team's performance in the quiz.

But here's the catch. In non-quiz situations, *everyone* has access to the internet. Every group can use the internet and other artificial enhancements. So, including internet-connected cyborgs on your team offers no marginal gains in cognitive diversity, it is simply standard practice. What *does* increase cognitive diversity compared to other groups is to have a member who can *use* the internet in a way that others cannot. The benefit of cognitive diversity from cyborgs in this case reduces, in comparative terms, to the benefit of cognitive diversity in humans.

Of course, in some situations complex technology can create cyborgs that offer considerable increases cognitive diversity compared to human only groups. A human with an exoskeleton robotic body may offer benefits of cognitive diversity, as may a human with cameras, recording equipment or other sensors. A submarine equipped with a sonar sensor, or a hospital with a radiology department increases their cognitive diversity in that they can access new types of evidence. A group with both a radiology department *and* a sonar can effectively study sea-life even better than a group with only one – so the cognitive diversity increases in the usual way. Just as having *two* group members with diverse epistemic standpoints increases cognitive diversity further than only have one diverse member.

So, what's the point? The idea here is that cyborgs, that is, human-artificial hybrid agents, offer increases in cognitive diversity compared to groups without hybrid agents. Their presence is, clearly (as we see from the universality of cyborg-inclusive groups) beneficial for the group. That is, there is a clear epistemic advantage to having various types of cyborgs in a group. And that advantage, I suggest, is well explained by the corresponding increase in cognitive diversity. This can be taken as evidence for the claim that cognitive diversity is beneficial for a group.

So artificial systems in general increase cognitive diversity and, in fact, it is apparent that this is one of the primary reasons for which we use artificial systems (why else, I wonder, do humans use sonar, the internet, or X-ray machines?). They are used, most of the time, as epistemic enhancers. This leads to two conclusions. First, the benefits of using cyborgs can be explained in terms of cognitive diversity, and those benefits are powerful and important. Second, defending premise (2) requires that advanced artificial systems not only offer *some*

cognitive diversity compared a human-only group, but also offers *distinct* cognitive diversity to cyborg-human groups. Otherwise, there is no rational motivation to develop advanced artificial systems, since existing cyborg-human groups would already be just as cognitively diverse. That is the issue I will now focus upon.

Here I argue that advanced artificial systems increase cognitive diversity in a distinct manner. For advanced artificial systems to do this, they must offer different (and useful) beliefs, aims, sources of evidence, etc. – in short, they must offer a distinct *epistemic standpoint*.

Artificial systems already offer many of these benefits through cyborgs. AI has recently had profound success in various areas such as image generation and conversational ability, while it has had long standing and dramatic benefits for computationally intensive tasks such as those required in economics, physics, and engineering. However, these types of abilities are already available to cyborgs. A cyborg increases cognitive diversity in a group compared to a human group without cyborgs, but, consequently, an advanced artificial system with equivalent abilities would not increase cognitive diversity more than a cyborg would.

So, to increase cognitive diversity in even cyborg groups, advanced artificial systems need to offer an epistemic standpoint that even cyborgs cannot take. I propose that they do. An advanced artificial system is, I will argue, likely to distinctly increase cognitive diversity because it has a different *ontology* to both humans and cyborgs. One analogy that can help in imagining this is between animals and humans – animals, to the extent that they adopt an epistemic standpoint at all[8], adopt a distinct range of epistemic standpoints to humans because of their different ontology[9].

Advanced artificial systems, because of their distinct ontology compared to cyborgs, may adopt a distinct epistemic standpoint of their own – indeed, if variation in ontological structure tracks variation in epistemic standpoint, then advanced artificial systems are likely to adopt the most radical epistemic standpoint out of any agent on Earth. I now discuss four (ontologically driven) features of advanced artificial systems that lead them to have a distinct

[8] Aliens with the required intellectual capacities to genuinely adopt an epistemic standpoint are a better example, but animals are easier to describe.

[9] Interestingly, animal-human groups have higher cognitive diversity than human-only groups (and biodiverse groups of animals have higher cognitive diversity than homogenous ones). Of course, the epistemic advantages are even more difficult to realise, but it is sometimes done, most often with dogs (for finding drugs and survivors, guiding blind humans, and guarding humans or sheep).

epistemic standpoint. First, they are free from biological drives, second, they can have different sensory modalities, third, they have qualitatively different cognitive architecture; fourth, they can *integrate* all these differences into a single decision-making process.

Advanced artificial systems may gain a unique epistemic standpoint through being non-biological. advanced artificial systems would avoid the bias of being a biological system[10]. Humans and animals are naturally bound to value food, shelter and water. advanced artificial systems would not have the same trappings (though it may value electricity and not being turned off in some sense). Similarly, advanced artificial systems, if embodied robotically, would not have the same types of biological drives, such as fears of social isolation, physical harm, or uncertainty, or drives for social status and reproduction.

But a different ontology has wider ranging effects on advanced artificial systems' epistemic standpoints than merely avoiding biological drives. advanced artificial systems ontology can support different sensory modalities to humans. This can be observed in the tools that we already use: an X-ray scanner can 'see' the world in a way that no human agent can. Just like animals sometimes have different or more advanced senses compared to humans. An advanced artificial system may access evidence about microscopic physical facts, or about UV radiation, or even the exact temperature of the room in a perception-like way. These types of ability cause advanced artificial systems to adopt a different epistemic standpoint: the things 'apparent' and 'taken for granted' by the advanced artificial systems will be different, from a physical, and consequently conceptual perspective, from the things that humans take for granted.

Of course, this is no additional benefit compared to a human group with an X-ray scanner. A radiology expert cyborg can 'read' or 'see' the information offered in a roughly equivalent way. They can have intuitions about the information involved, for example, and understand subtle differences non-experts are unable to. They may 'take for granted' the same types of physical facts that an advanced artificial system with X-ray vision does.

However, advanced artificial systems offer a *greater* increase in cognitive diversity compared to the radiology cyborg for two reasons. The first reason is that advanced artificial systems' sensory modality is *fully integrated* into their epistemology. A bat can use echolocation in a

[10] Unless developed through synthetic biology.

dramatically more intuitive and expert way than a human with an echolocation device because the bat is *fully integrated* with echolocation – it cannot exist *outside* the echolocation-informed epistemic space. An advanced artificial system with an integrated, say, X-ray scanner or thermal vision, may be similar – and there may be corresponding increases in their ability to use the evidence involved. This is particularly useful in time sensitive situations. Imagine, for example, an advanced artificial system with various visual spectrums might be in aiding emergency medical diagnosis – they would be much more valuable, given the time constraints, than a human that uses X-ray and MRI scanners.

There is a second reason in the diagnosis example – one of the distinct advantages of an advanced artificial systems may be the *combination* of different sensory modalities that a human/cyborg is simply unable (cognitively) to do. A radiology cyborg may be able to intuitively understand X-rays, but they cannot simultaneously understand the full spectrum of radiation. Who knows, we might wonder, what kind of correlations might be derived from a fuller epistemic picture, such as the one a full radiation spectrum perceiving advanced artificial systems provides. Their expertise would cover a much greater range of evidence, and subsequently the inferences (or intuitions) can be expected to be significantly more accurate or useful.

An advanced artificial system would also have a qualitatively different cognitive architecture to humans. The differences are widespread, but one accessible example is memory. Unlike human biological memory, artificial 'memory' is not limited by capacity to represent information, but by the speed at which the information can be accessed. This means that an advanced artificial system's decision-making process, methodology, and conceptual frameworks can have a different texture compared to humans.

Of course, human memory, as well as human cognitive processing, is valuable precisely because it can do a lot with relatively compressed information, so advanced artificial systems knowledge would not be a replacement or straight upgrade on human knowledge and memory, but, in line with the benefits of cognitive diversity, would nonetheless improve the decision-making performance of their group.

Once again, this is particularly true for advanced artificial systems rather than internet-connected or database-wielding cyborgs. An advanced artificial system would be able to select the information it deems relevant quickly and independently. Unlike a cyborg who must trawl through a database. We can already see some of these abilities in large language

models like ChatGPT, who can present relevant information on a wide range of information much faster than a human with access to the internet.

Though it should be emphasised that this is unlikely to be an improvement in *speed* alone, but rather a *qualitative difference*; an advanced artificial system could present *different* reasons, rather than the same reasons faster. Perhaps the potential depth of its memory would allow it to present overlooked evidence, or perhaps it would allow it to generate complex statistical trends that a human could not perceive. Both capacities are present in cyborgs, who can perform statistical analysis and deep searches. But an advanced artificial system would be able to combine these abilities with autonomous memory access in a coherent and independent way that would, I suggest, lead to independent and innovative contributions to decisions.

Finally, the most important feature of advanced artificial systems in increasing cognitive diversity would be in their decision-making process and cognitive architecture. This has been hinted at in the previous paragraphs, but the integration of these abilities into the independent decision-making structure of advanced artificial systems would be an essential factor to their increasing cognitive diversity. An advanced artificial system would increase cognitive diversity because of the emergent capacities of being able to use all the computational tools available to it – combining artificial epistemic capacities, the lack of biological trappings, and artificial memory (in the form of databases and so on) with the ability to unite all of these into coherent, relevant pieces of evidence. This integrative capacity[11] is the essential factor in advanced artificial systems contributing to cognitive diversity more than humans equipped with a combination of ultimately similarly functioning humans with AI tools. It is also one of the biggest challenges in developing advanced artificial systems. Fortunately for this chapter, I do not have to suggest a solution for this, as I am arguing that we have a reason to develop advanced artificial systems (and perhaps not even a definitive one), not that we *can* do so.

That said, if integrated, holistic cognitive architecture is necessary for the most profound benefits of advanced artificial systems, then the possibility of an advanced artificial systems without this would undermine the argument. While my point here is necessarily somewhat speculative, it seems to me improbable that an advanced artificial system could lack this type

[11]To me, there are parallels with Kant's ideas of the unity of apperception, and cognitive science's holistic theories of mind here, but pursuing those parallels would be diversionary.

of cognitive architecture because achieving human level functionality will probably involve the ability to evaluate decisions and beliefs with respect to the agent's total epistemic situation. Of course, oversights are permitted (and common in human thinking), but in principle humans can integrate all the evidence at their disposal in their decision-making processes. We ought to expect that an advanced artificial system will do the same. This does not, it should be noted, imply that advanced artificial systems will necessarily have any kind of phenomenological experience – or even that an advanced artificial systems' individual functions must be product of the *same* cognitive process. Instead, what is needed is a high-level of communication between different cognitive mechanisms, or an 'executive function' capable of uniting individual functions and sources of evidence into a single coherent explanation.

There are two final points for defending premise (2). First, advanced artificial systems may contribute to groups' decisions in relatively unforeseen ways, such as by offering fresh and interesting *moral, aesthetic* or *philosophical* analogies, evidence, and reasoning. I will not go into the full details of that here, but the contribution of advanced artificial systems to cognitive diversity should be, in theory, a general contribution that improves *all types* of decision making – including those that are very difficult, complex, or unusual. Second, I make no claims here that advanced artificial systems will be *better* in any particular or general sense than any human or able to make *better decisions* than any human. The point here is for cognitive *diversity,* which can be promoted by systems that make universally *worse* decisions (such as using a child's analogies to stimulate a group's imagination).

The sum of these differences, which are somewhat speculative, as all things about potential futures must be, is that advanced artificial systems would almost certainly have a *radically* distinct epistemic standpoint compared to humans, and even compared to cyborgs. Thus, they would have a correspondingly (we can assume) large epistemic advantage to offer to a group of which they are a member. Thus, premise (2) is true.

Some might think that the epistemic advantage advanced artificial systems may offer will be very challenging for a human/cyborg group to take. One piece of evidence for this, as the previous section suggested, is that cognitive diversity in human groups does not lead to corresponding increases in performance because it is correlated with an increase in group conflict. advanced artificial systems may also increase group conflict. Most people do not like being told what to do by a machine (although we get used to it surprisingly quickly).

One line of response is that the increase in cognitive diversity is *so* large that it would still be an overall positive reason to develop advanced artificial systems even if they did increase group conflict. Furthermore, humans have fewer biases towards artificial agents and often work quite well with them. Human/advanced artificial systems groups will have fewer clashes of biological drives, and less competition for status or validation. These features may reduce group conflict. Of course, depending on the humans (or cyborgs) in the group, they may just as well increase it.

2.7. Conclusion

I have argued that premises (1) is likely true. Cognitive diversity is beneficial to a group's decision making, and this can be seen from several perspectives. Cognitive diversity appears to improve a group's collective ability to make decisions compared to a homogenous group. Experimental evidence from philosophy of science supports this and is an important subject of attention in that field; while there is theoretical support from epistemology, wherein diverse perspectives that are appropriately integrated into a decision-making process improve the accuracy and robustness of the outputted decisions and theories.

Premise (2) is more complex, partly because it is difficult to predict the nature of advanced artificial systems even should we be able to develop it. I have tried to make minimal assumptions about the nature of advanced artificial systems, assuming that it will be nothing more than an integrated composite of abilities that our artificial systems already have. If advanced artificial systems have this constitution, then there are many reasons (and I have only covered a few) to believe that they would increase cognitive diversity in a group in which they are present. If an advanced artificial system can be appropriately used as part of a groups decision making process (which seems to me entirely reasonable) then it would *beneficially* increase cognitive diversity.

The truth of the cognitive diversity reason depends on the truth of (1) and (2). I have argued that both are true. If this is correct, then our decision about whether it is permissible, beneficial, or obligatory to develop advanced artificial systems ought to consider cognitive diversity reason as a contributing reason. As noted in the introduction, there is no need for this to be a definitive or overwhelming reason to develop advanced artificial systems, but it is yet an undefeated reason that ought to contribute to our decision.

3. Conditions for Moral Agency

3.1. Introduction

There are many reasons to want to develop artificial moral agents. I presented one reason that I think is especially powerful: advanced artificial systems like artificial moral agents may improve our theories, making them more objective and explanatorily effective. That reason does not, I admitted, decisively warrant investing serious resources into developing artificial moral agents. Nonetheless, I leave the overall justification of the project of machine ethics with the following point, which I think vindicates, at least, this thesis. Perhaps building an artificial moral agent is wrong; but even so, it is beneficial to know if and how they could be built. This is the knowledge that this thesis aims to progress towards – even as a hypothetical case or theoretical example, explorations of the possibility and prospects of artificial moral agency can offer many contributions to philosophical thought.

So, with the question of motivation behind us, I turn to a foundational concept of machine ethics: moral agency. Clearly, in exploring whether artificial moral agents are possible (and, assuming they are, exploring *how* you could make one), it is useful to know what exactly it is that we are evaluating. Moral agency has no universally agreed upon necessary and sufficient conditions. I here consider various theories of moral agency and settle on what is intended to be a wide-ranging and uncontroversial definition of the concept.

This chapter offers two necessary conditions of moral agency. The following section (3.2) offers a brief overview of concept of 'agency' being used. I use a moderate definition of agency under which artificial systems are unproblematically agents. Section 3.3. offers an overview of the concept of moral agency. I argue to be a moral agent an agent needs two capacities. 3.4. argues that moral agency requires an agent have the capacity to be responsive to moral reasons. 3.5. outlines the idea that moral agency requires the capacity to be responsible. While the condition of the capacity to be responsible is reasonable and supported from several directions, there are some who deny this condition. But I defend it as an implication of the standard understanding of moral agency.

3.2. Agency

It is standard in discussions of artificial systems to discuss whether they are agents. In our everyday language, 'agency' has two forms. First, humans are the standard agent – they have intentional states (i.e., beliefs, desires, motivations, and so on); and they interact with the environment in a way that reflects this. Second, we sometimes use a broader sense of 'agent' to mean any system that performs actions – for example, a sportsperson or author has an 'agent' who acts on their behalf; software engineers program 'agents' to carry out tasks; and the army or police deploy 'agents' to particular locations. I am interested in the first definition of agency here. Focusing on this type of agency does not typically cause confusion in philosophical literature, but since the concern here is *artificial systems*, and some artificial systems are referred to as 'agents' in the second sense, it is a distinction worth making.

One approach that I do not use is to develop a multi-tiered 'scale' of agency (see Moor, 2006). I use this approach for neither agency nor moral agency. In my view, agency (and moral agency) is a binary concept. One can either be an agent or not an agent. Similarly, I do not, as some do in philosophy of action, equate 'agency' with 'autonomous agency' or 'responsible agency' – for me an agent need not possess any properties or attributes that humans might routinely fail to possess. Here it is assumed that humans are our paradigmatic exemplar of agency, and the average human is an indisputably an agent, I'll discuss the reason for this shortly.

Having described what I'm *not* doing, I now turn to what I *am* doing – building a sketch of an account of agency that can support the work in this thesis. The account I develop here is not, I should emphasise, intended to be precise or well-defended. It is intended to be a broad foundation from which to consider whether artificial systems can be moral agents without being distracted by whether they might be said to be 'agents' proper. There are two accounts of 'agency' that I am working from. Christian List & Philip Pettit (2011) claim that an 'agent' has three features:

> First feature. It has representational states that depict how things are in the environment.
>
> Second feature. It has motivational states that specify how it requires things to be in the environment.

> Third feature. It has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification.
>
> An 'agent', on our account, is a system with these features: it has representational states, motivational states, and a capacity to process them and to act on their basis. When processed appropriately, the representational states co-vary with certain variations in the environment […]. And the motivational states leave the agent at rest or trigger action, depending on whether the motivating specifications are realized or unrealized in the represented environment. (List & Pettit, 2011, p. 20)

Luciano Floridi and J.W. Sanders' (2004) offer a more substantial definition while avoiding using intentional notions such as 'motivational'. They use the example of 'Jan', claiming that "Jan is an agent if Jan is a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it…" and possesses interactivity, autonomy and adaptability, which are themselves explained:

> a) Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient – for example gravitational force between bodies.
>
> (b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.
>
> (c) Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed […] as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state [...] then adaptability follows from the other two conditions. (Floridi & Sanders, 2004, 357-8)

While both are general accounts of agency (and there are surely others), both are intended to apply to artificial agents. List and Pettit's is in the context of group agency, but they use a robot as their example case for generating the definition, and groups and artificial agents are similar in being possible non-human agents. Floridi & Sanders' is in the context of artificial agency. There is a lot of overlap between them, and both hold that artificial systems can be agents.

On both accounts, the central aspect of agency is the production of actions through mediating rules and representations that are 'internal' to it in the relevant respect[12]. So, things like weather systems, rocks, natural forces, and hurricanes[13] are not agents, because they do not have internal representations and motivations (for List and Pettit) or internal state transitions and the ability to change their own transition rules (for Floridi & Sanders). While things like humans, animals (including simple ones like bees and slugs) and (complex) artificial systems are agents.

For examples of artificial systems specifically, under both accounts a thermometer is not an agent (nor a hammer nor a tape recorder). But an artificial system that uses representations to achieve its motivations through action or has internal rules that can be adjusted to suit different situations is an agent. So, word processors (motivated to spellcheck this chapter, and changes its internal rules to reflect feedback and aims); robot vacuum cleaners (motivated to clean the room, can represent a map of the room internally, can adapt its behaviour to fit that representation); and chatbots (motivated to produce content, can adapt to the prompts given and feedback on those prompts, can represent the text of others and itself) are all agents. They are artificial agents.

One primary difference between these accounts is that Floridi & Sanders' account avoids intentional state terminology – that is, it holds that 'motivations', 'beliefs', 'desires' and 'intentions' and so on are unnecessary for agency[14]. List and Pettit's account, in contrast, holds that agents need to have intentional states. This is the mainstream position, which is also taken by popular 'belief-desire-intention' and 'taking as a reason' models of agency (Schlosser, 2019). Kenneth Einar Himma, for example, takes intentional states to be necessary for agency: "X is an agent if and only if X can instantiate intentional mental states capable of directly causing a performance" (2009).

This leads to an obvious question that I will leave open for now: in what sense is attributing intentional states to artificial systems appropriate? If it is, then is it done 'metaphorically' or

[12] Having 'internal representations' is nice and easy for robots and humans but is less obvious for decentralised systems like collectives, distributed artificial systems, hybrid systems. I assume that these decentralised systems do have internal representations in some sense.
[13] Some conceptions of God would be agents under this account – but many of the properties of God, such as not being embodied, would seem to preclude it from agency.
[14] It is also relatively representation-free, though that depends on one's understanding of representation & mental representation.

'literally'? It is widely held and has been convincingly argued by Daniel Dennett (1981) and other 'instrumentalists' that it is appropriate, in some sense, to attribute intentional states to artificial systems. It is more controversial whether this attribution is anything more than 'metaphorical', that is, of a (perhaps radically) different kind to the 'literal' attribution of intentional states to humans.

Chapter 5 argues that artificial systems should be attributed intentional states in a literal and non-metaphorical way. But here I adopt a moderate model of agency that only requires possession of intentional states in an instrumentalist and metaphorical way that is satisfied by artificial systems. The account of agency being used here, then, is closer to the List & Pettit model of agency than the Floridi & Sanders model. For the following sections, then, I will describe artificial systems as agents with intentional states.

With this, I can outline the conditions of agency that I endorse, as a combination of the accounts just discussed, are:

a) An agent must be situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it.
b) An agent and its environment (can) act upon each other.
c) An agent is able to change state without direct response to interaction: it can perform internal transitions to change its state.
d) An agent's interactions (can) change the transition rules by which it changes state.
e) An agent has representational states that depict how things are in the environment.
f) An agent has motivational states that specify how it requires things to be in the environment.
g) An agent has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification.

This is a lot of conditions, but they do not amount to a significantly constraining definition of agency. They mostly serve to eliminate candidates for agency that are clearly not agents, like rocks, hurricanes, weather systems and so on. On this account, most animals are agents, and I suspect that many simple organisms like bacteria and fungi are too. Many complex artificial systems are agents, while many simpler artificial systems are not agents.

One further point is that being an agent on this account means being responsive to reasons to some extent (Fischer & Ravizza, 1998). Agents are agents in virtue of their ability to change the environment to achieve some goal-state or satisfy a motivation. And this is tantamount to being responsive to reasons – a goal or motivation is a 'reason' to change the environment. Similarly, in a complex environment an agent may only be able to satisfy some goals at the expense of other goals. Navigating this kind of situation requires an agent to 'balance' their motivations in a process that can easily be understood as deliberative reasoning – comparing the strengths of one reason with another. Agents, then, are (minimally) reasons-responsive.

To 'respond' to a reason is to act in accordance with that reason. This means that an agent acts in a way that is best understood, given their knowledge, beliefs, and aims, as them acting *for* the reason. Two things come out of this, first, responsiveness to reasons here is *normative* in the sense that it is a 'good', factual reason for acting rather than simply an explanation for their behaviour. This can be distinguished from a 'motivating' reason, as Herman Veluwenkamp explains:

> Let us look at an example. Alida might think that it is going to rain, and therefore bring an umbrella to work. We can say that Alida has a motivating reason to bring an umbrella. The motivating reason is the belief that it is going to rain and her desire not to get wet. But this doesn't mean that Alida also has a normative reason to bring an umbrella to work. Of course, if Alida has a motivating reason to bring an umbrella, then so takes herself to have a normative reason as well. There are, however, two different ways in which Alida can be mistaken about this. Firstly, she can make a normative mistake: she can fail to realise that the fact that it is going to rain is a reason to bring an umbrella. And, secondly, she can make a non-normative mistake: she can be mistaken about the weather. In both cases, Alida has a motivating reason without the accompanying normative reason. (Veluwenkamp, 2022)

Agents do, of course, have both motivating and normative reasons-responsiveness. But motivating reasons-responsiveness is not sufficient for agency. One needs normative reasons-responsiveness in the sense that their motivations reflect *good* reasons to act. If a system cannot respond to normative reasons, then they cannot intervene suitably in the environment. While agents may not always have motivational reasons that track their normative reasons, to be an agent a system does need some of their motivational reasons to also be normative reasons. Otherwise it would be wrong to interpret them, even in the loose instrumentalist sense that I am using, as having intentional states.

Even with intentional state attribution, reasons-responsiveness, and agency, there are still additional conditions that need to be met for an agent to be a moral agent. So, while on the

account of agency used here artificial systems can be agents, the possibility remains that they cannot be *moral agents.*

3.3. Moral Agency

Moral Agents are agents that meet some additional criteria. Adult humans are the paradigmatic moral agent[15], but artificial systems and animals[16] seem to be borderline cases. In the previous section I outlined the account of agency I will use. Under this account artificial systems can be agents. A natural follow-up question is whether they can be moral agents, too. However, the nature of moral agency is subject to some debate, and some argue that artificial systems, whether they are agents or not, cannot be moral agents because they will never be able to possess the necessary capacities.

Raul Hakli and Pekka Mäkelä point out how moral agency (unlike agency simpliciter) is often taken to be centred on distinctly human and hard to quantify capacities:

> "Even though it has been difficult to lay out the necessary and sufficient conditions of moral agency, people have traditionally thought of themselves as moral agents. They have thought that adult human beings who have normal abilities and who have gone through typical processes of upbringing and socialization qualify as agents that can be appropriately held morally responsible for their intentional actions and behavior. The idea has been that whatever the exact conditions of moral agency are, they should be within reach for typical adult human beings. As a result, suggested capacities required by moral agency have closely resembled typical capacities of adult human beings, capacities relating to intentionality, rationality, sentience, autonomy, normative understanding, sociality, and personhood." (Hakli and Mäkelä, 2019).

If moral agency requires distinctly human capacities, the prospects for artificial moral agency look slim. Some defend these distinctively human conditions, and claim that an agent needs consciousness, personhood, and the like, to be a moral agent. But so-called 'functionalist' accounts of moral agency offer different conditions of moral agency, ones more easily met by artificial systems.

---

[15] Like they are for agency. Though the concepts are clearly distinct. Humans are versatile paradigms.
[16] See Rowlands, 2012, for an example of animal moral agency discussion.

One strategy for finding the precise conditions of moral agency is to try to work to those conditions from popular intuitions about individual cases. Hakli and Mäkelä demonstrate this kind of reasoning:

> "As most people, many philosophers included, seem to share a rather firm intuition that robots and AI agents are not moral agents, the situation has created a need to sharpen the conditions of moral agency in a way that would still include adult human beings but exclude these artificial agents." (Hakli and Mäkelä, 2019).

I agree that it *seems* true that contemporary artificial agents are not moral agents. Even if it *were* true, that bears little impact on whether artificial agents *can be* moral agents. But, putting that aside, I want to draw attention to the argumentative strategy being used here.

The strategy of taking popular intuitions about cases and defining the conditions of moral agency *such that* our intuitions are correct is flawed. Some reliance on intuitions about cases is inevitable in theory-making, but it is not, and should not be, surprising when philosophical or scientific theories run contrary to prevailing intuitions. Quantum mechanics and general relativity, for example, are profoundly unintuitive to many, but that ought not be taken as serious evidence that they are wrong. The concept of moral agency demarcated by our intuitions is our 'folk' concept of moral agency – and the best theory of moral agency might depart from that.

But that is not to say that folk conceptions are useless or that intuitions lack value. Radical departure from a folk concept is only warranted with strong evidence that the folk concept is based on misconceptions or confusions. One example: a tomato is a fruit because the theory of fruit is well-supported by our other theories and evidence, despite the common pre-theoretical intuition that tomatoes are not fruit. But a theory of fruit which says that meat is fruit is not similarly well-supported, partly because our pre-theoretical intuitions about meat not being fruit are so strong. Simply, a good theory ought to be *informed* by the relevant folk notions without being fashioned entirely in its image. A theory of moral agency, then, should only depart from pre-theoretical intuitions with good reason. Most of the time, that good reason is that the theory can gain significant advantages – advantages such as better explaining evidence, cohering with other well-supported theories, and being logically consistent.

I suggest that one intuition about the concept of moral agency ought to be accepted as a central and foundational conceptual necessity. This is that the average, competent, adult human is a moral agent. Any theory of moral agency that entails that most humans are not

moral agents departs from the concept of moral agency. I do not make this assumption about moral agency out of rampant anthropocentrism, but because our concept of moral agency was always intended to and normally is used to describe humans. Moral agency need not, by that token, be *restricted* to humans, but humans are, at least for now, the prototype moral agent. Just as it would be unwise to throw out the International Prototype Kilogram before agreeing on the precise physical characteristics of a kilogram, it is unwise to throw out the moral agency prototype adult humans before agreeing on the precise conditions of moral agency.

A second intuition that I think fundamental is that it is possible, in principle, for a non-human agent to be a moral agent. That is, it is not conceptually necessary that only humans can be moral agents. The concept of moral agency is a set of conditions that can denote *any* agent that satisfies them. Assuming otherwise seems anthropocentric and has been argued against (See Coeckelbergh, 2012; Gunkel, 2012). I discuss this further in 4.6.

In the following two sections (3.4. & 3.5.), I defend two necessary conditions of moral agency. To be a moral agent, I will argue, an agent must have the capacity to adequately respond to moral reasons and the capacity to be responsible. These conditions are neither precise nor necessarily complete, so I will retain the services of the conceptual prototype - adult humans – as a measuring stick.

3.4. The Capacity to Respond to Moral Reasons

I will argue that moral agency is generally accepted to require the capacity to respond to moral reasons (or 'moral reasons-responsiveness'). Per 3.2., an agent is, by definition, responsive to reasons. One natural (and, I will argue, correct) move is to suggest that *moral* agents are distinct from agents partly because they are responsive to *moral* reasons[17]. I will now offer some clarifications about this condition.

First off, the proposed condition is that moral agency requires the *capacity* to respond to moral reasons and need not *actually* respond to moral reasons. Moral agency only requires the *capacity* to respond to moral reasons because of the following cases. First, imagine a wily

[17] Some use 'moral rationality', perhaps 'moral competence', or having 'moral psychology' to refer to the same underlying capacity.

moral agent refuses to respond to *any* moral reasons while having a perfectly good capacity to do so. If they do this, the agent ought not be able to shed their moral agency. Moral agency is not the sort of thing that an agent can opt out of that easily. Second, imagine an agent is prevented from responding to moral reasons through physical constraint or coercion while retaining the *capacity*[18] to respond to moral reasons without *actually* responding to them. Once again, these agents still seem to be moral agents despite not responding to moral reasons.

One implication of this, which I endorse and take to be uncontroversial, is that moral agency is a *status*. Moreover, it is a *resilient* status: it cannot easily be shed and does not vary with the normal procession of reasoning and action. Simply, one cannot be a moral agent for some actions and the next minute, without significant changes, fail to be a moral agent for others. That is not to say, however, that moral agency applies to an agent throughout their lifespan. Moral agency is a resilient but not inalienable status. Children are not moral agents, but they *become* moral agents at a certain point, and then continue to be moral agents so long as there are no significant changes in their capacities. Adult humans can cease to be moral agents by losing a necessary capacity for moral agency – death obviously puts a permanent end to moral agency, but things like dementia, comas, brain damage and some types of brainwashing can cause a moral agent to shed (perhaps temporarily) their status.

Now to the meat of the matter. What distinguishes a moral reason from any other type of reason? The domain of moral reasons can be contrasted with the domains of 'prudential', 'instrumental' or 'practical' reasons (Crisp, 2018). Moral reasons typically concern the interests of others - if you help another because you are motivated by some concern for them, you are responding to a moral reason; but if you help them because they are annoying and you want them to stop asking, you are not. Another way of putting it is that moral reasons are reasons most aptly described in moral terms, such as 'right', 'wrong', 'permissible', 'impermissible'. Of course, most reasons can be described in these terms – but moral reasons are those that are *essentially* or *centrally* described by them. Murder is centrally *wrong*, desiring to help another out of concern is centrally *right*; helping another motivated by annoyance might be *wrong* or *right* in some senses, but it is not *centrally* wrong.

---

[18] This unveils the counterfactual nature of a capacity – though I do not offer a full explanation of it. A crude explanation is that a capacity is a process that can produce an effect unless otherwise prevented from doing so.

I want to head off any metaethical objections: I am not interested in the nature of moral 'facts' – that is, whether moral facts are objective or mind-dependent, natural or non-natural, universal or particular, and so on. Moral agents, I contend, must be responsive to moral reasons *whatever* the metaethical status of moral facts turns out to be. There are some brief metaethical concerns I can offer a quick reply to. The first concern stems from the possibility that moral facts are so complex and/or idealised that no agent could possibly respond to them – that is, the possibility that 'moral error theory' is true. If so, a human taking themselves to be responding to a moral reason is not responding to anything moral at all. Second, consider the possibility that a simple hedonic utilitarianism is true and that moral facts *only* pertain to pleasure or pain – in this case some reasons, such as reasons for keeping promises, may not be moral reasons. Other moral theories may lead to similar consequences if they turn out true: if consequentialism is true perhaps some of deontology's moral reasons are not really moral, and vice versa.

Do these types of concern spell trouble for the condition of moral agency being proposed? I think not. Moral reasons as understood here refer to a domain of reasons, but that domain ought not be restricted to reasons that always successfully reflect moral facts. Whatever metaethical theory is true, people still act for moral reasons in this sense. That a reason *could* reflect a moral fact is probably sufficient for it being a 'moral reason' in the sense of 'within the domain of morality'.[19] Morality is here understood, then, in a descriptive sense – a moral reason is a reason that plays a role in a plausible moral theory. It is not being understood in the normative or deeply metaphysical sense, in which moral reasons are the reasons that the true moral theory concerns.

The capacity to respond to moral reasons is not here taken to necessarily involve an agent being able to explain *which* moral reason they are responding to[20]. For example, a person without the ability to speak may still respond to moral reasons despite being unable to explain them – this can be true even if they have no conceptual ability at all to understand, in

[19] One worry might be that this opens up the possibility of quirky moral reasons. For example, if a group of people believed that all moral decisions ought to be determined by the flip of a coin. They are not, it seems, responsive to moral reasons, but you might think that coin-flip reasons *could* reflect moral facts. I have no entirely satisfying explanation for deciding which reasons *could* reflect moral facts. I only suggest that the moral theory which takes coin-flips as moral reasons is implausible, and that folk intuitions about such plausibility are generally sufficient for distinguishing between moral and non-moral reasons in this way.

[20] Or even to be able to explain *that* they are responding to a moral reason at all.

propositional terms, the reason they are responding to. Imagine, for example, a person unable to speak who is entirely motivated by, say sympathetic emotions or an unreflective but refined process of mimicking others. Suppose that through this process they successfully respond to moral reasons. The responsiveness to moral reasons (and responsiveness to reasons in general) envisaged here does not, therefore, require the agent to have rich internal representations or conceptual understanding of that reason. Though of course you, reader, can take your own mental representations of moral reasons as an indication that you have the *capacity* to respond to them. But without a corresponding action or utterance that representation does not *demonstrate* that capacity. If an agent demonstrates the capacity, it no longer matters whether they have subjective evidence that they possess it. So, rich internal representations, conceptual understanding, and the ability to explain the moral reason, are neither necessary nor sufficient for responding to moral reasons.

Finally, moral agency does not require the capacity to respond to *every* moral reason. But it requires the capacity to respond to *moral reasons* as a general class of reason – I call this 'adequate' moral reasons-responsiveness. Presumably, if you have the capacity to respond to one moral reason, then you have the capacity to respond to many (This is the view of Fischer & Ravizza, 1998). But a hypothetical agent with the capacity to respond only to a single or very limited set of moral reasons would not *adequately* respond to moral reasons. One measure for adequacy for responding to moral reasons is *human equivalence*. On this measure, an agent that can respond to moral reasons in a roughly performatively equivalent way to typical humans has adequate moral reasons-responsiveness. Some think the bar of adequacy is lower than human-equivalence (e.g., Bekoff & Pierce, 2010; for further examples, see Clement, 2013), but human-equivalence is the lowest bar that is uncontroversial.

Another relevant area is research into the cognitive mechanisms that humans respond to moral reasons with – that is, 'moral psychology'. Is the human capacity to respond to moral reasons performed using sentiments, intuition, reasoning, or something else entirely? This issue is important and will be discussed in chapter 8. But it is not relevant to the moral agency condition –*some* cognitive mechanism is needed to respond to moral reasons, but I leave the exact nature of it open.

To summarise, I have outlined the idea that moral agency requires the capacity to respond to moral reasons. This is intended to be metaethically neutral. Moral reasons are the domain of

reasons that can be part of a plausible moral theory. To be a moral agent, an agent needs to have the capacity to *adequately* respond to moral reasons – that is, they must be able to respond to a sufficient (I assume this is human-equivalent, but the condition of moral agency can accommodate varying bars of adequacy) range of moral reasons.

The condition of moral agency I propose is thus:

(a) To be a moral agent, an agent must have the capacity to be adequately responsive to moral reasons.

Given the clarifications above, this condition is uncontroversial. Almost every account of moral agency accepts it, in some form or other (i.e., while using different terminology). Some examples follow.

List (2021) defines a moral agent as "an agent with the capacity to make normative judgements about its choices — judgements about what is right and wrong, permissible and impermissible — and to respond appropriately to those judgements". Many other accounts contend that the ability to make moral judgements is necessary for moral agency (for example, Purves et al., 2015, who reappear in chapter 6.). Moral judgement is one way of demonstrating or achieving adequate moral reasons-responsiveness. It is perhaps more than adequate moral reasons-responsiveness, in that moral judgements may imply other capabilities or processes that adequate moral reasons-responsiveness does not. But those who think that moral judgement is necessary for moral agency must, I suggest, agree that moral agents must be adequately moral reasons-responsive.

The same applies for the popular view that moral agency requires moral sentiments or the ability to empathise (Kauppinen, 2022). Again, moral sentiments can be plausibly said to be one means of responding to moral reasons. And, again, moral reasons-responsiveness is a necessary but not sufficient function of moral sentiments. This is generally taken for granted for both moral sentiments and moral judgement – I know no-one who argues that one can have the capacities for moral judgement or moral sentiments necessary for moral agency *without* being able to respond to moral reasons. Indeed, it would be absurd to do so, as responding to moral reasons is *central* and *essential* to these capacities. Likewise, if one had

the capacity for moral judgement[21] or moral sentiments *without* being *adequately* moral reasons-responsive, I think that it would be agreed that they are not moral agents. Both, therefore, accept condition (a) as a necessary condition of moral agency.

Floridi & Sanders (2004) say that an agent is a moral agent when it is in a 'morally charged' context. Despite the different terminology (i.e., aversion to intentional terms) of their account, they seem to fundamentally agree that a moral agent is distinguished from ordinary agents by being responsive to moral reasons. Being put in a 'morally charged' situation is to be put in contact with moral reasons. Floridi & Sanders place the bar of adequacy much lower than human equivalence, however. Thinking that respond to few or even only one moral reason is sufficient for moral agency.

3.5. The Capacity to Be Responsible

The second condition of moral agency I propose is that to be a moral agent an agent must have the capacity to be responsible. There are various stances on the relationship between responsibility and moral agency.

Dorna Behdadi and Christian Munthe (2020) claim that a section of the literature "assumes that moral agency requires moral responsibility". We can see this kind of implication in the passage from Hakli and Mäkelä in 3.3, who suggest that moral agents are "agents that can be appropriately held morally responsible". Behdadi and Munthe contrast these positions with the opposing position that "moral responsibility requires moral agency". However, I will argue that those taken to assume that moral agency requires moral responsibility are really assuming that moral agency requires the *capacity* to be morally responsible. Thus, according to these accounts, one does not need to *be* responsible to be a moral agent.

Before defending this, we should be armed with a rough understanding of what responsibility is. Responsibility is typically understood to concern 'control' and 'epistemic' conditions. That is, to be responsible for an action an agent must have adequate 'control' over the action and satisfy 'epistemic' conditions concerning their ability to form beliefs about the relevant

[21] Which may be possible if moral judgement does not necessarily entail adequate moral reasons-responsiveness, which it is often taken to do. Perhaps 'malfunctioning' moral judgement could lead to inadequate moral reasons-responsiveness.

facts. Epistemic conditions are normally easy to meet – they are intended to show that an agent cannot be responsible for consequences of their actions that they could not reasonably be expected to anticipate. 'Control' conditions are more challenging – some think that to meet these conditions an agent needs to be autonomous or have free will (Frankfurt, 1988; Haji, 1998; Mele, 1995); and sometimes meeting control conditions are described in terms of reasons-responsiveness (Fischer & Ravizza, 1998; Kane, 1998; Wolf, 1993). Reasons-responsive control conditions for responsibility may *imply* adequate moral reasons-responsiveness described above; but targets a *generally* competent reasons-responsiveness.

Responsibility is often understood in a Strawsonian sense in which you are responsible if it is appropriate to blame or praise you. There are many disagreements about these appropriate responses to responsible actions (known as reactive attitudes). For my purposes it is not especially important – let us take the general line that being aptly subject to reactive attitudes is a good indication of responsibility. Thus, if one is responsible, one is generally blameworthy and praiseworthy for their actions.

Finally, some distinguish between 'accountability' and other types of responsibility (Shoemaker, 2015; Watson, 1996) – often Shoemaker's tripartite definition of responsibility as involving 'accountability', 'answerability' and 'attributability' is used. For my purposes, once more this is not a primary concern. The thesis generally refers to the capacity for responsibility, though a natural focus is on accountability – that is, the ability to change outcomes. Evidently, there is extensive literature on responsibility, and some of it will be discussed further in chapter 7. but as a rough outline of responsibility this should be sufficient.

Let us review the artificial moral agency literature that seems to argue that moral agency requires that an agent *is* responsible. First, Himma argues that responsibility is a central part of the standard conception of moral agency:

> The idea is that, as a conceptual matter, the behavior of a moral agent is governed by moral standards, while the behavior of something that is not a moral agent is not governed by moral standards. … To say that one's behavior is governed by moral standards and hence that one has moral duties or moral obligations is to say that one's behavior should be guided by and hence evaluated under those standards. Something subject to moral standards is accountable (or morally responsible) for its behavior under those standards. (Himma, 2009, 21).

Himma, (Along with several others such as Coeckelbergh, 2014; Sparrow, 2007; Torrance, 2014) thinks that being responsible has a strong conceptual relationship with moral agency. Being a moral agent is being subject to moral standards, and if you are subject to moral standards then you are- responsible for your behaviour. So, the temptation to interpret the position as positing an identity between responsibility and moral agency is understandable.

Some even think that responsibility is sufficient for moral agency. Carissa Véliz, for example, says that: "To be a moral agent just means that one is responsible for one's moral actions." (2021, 493). Mark Rowlands thinks that "X is a moral agent if and only if X is (a) morally responsible for, and so can be (b) morally evaluated (praised or blamed, broadly understood) for, its motives and actions." (Rowlands, 2012, Chapter 3)

Despite superficial appearances, all these accounts plausibly only argue that moral agency requires the *capacity* to be responsible. Himma and Hakli & Mäkelä can be read in this way – a moral agent *can* be responsible, but one does not need to be responsible to be a moral agent. This can even extend to Véliz and Rowlands, in which being responsible is dependent on a moral agent *performing moral actions* in (presumably) the appropriate, responsibility supporting conditions, and thus moral agency involves the capacity to be responsible.

Recall that moral agency is a resilient status. Is 'responsibility' likewise a resilient status? Or even a status at all? Unlike for 'moral agency', it is unclear if 'being responsible' is a status. Responsibility can be understood as episodic or dispositional. For example, when I say, "he is a responsible person", I am (normally) referring to dispositional, rather than episodic, responsibility. Whereas when I say, 'he is responsible for protecting the welfare of others', I refer to his episodic responsibility for a certain action or consequence. If responsibility is dispositional, then it is a status (and a fairly resilient one); if it is episodic, then, if it is a status at all it is a transient one, i.e., non-resilient or easily sheddable.

So, which type of responsibility is necessary for moral agency? It is reasonable to think that dispositional responsibility is what is required. Since moral agency is a resilient status, then it's conditions must also be resilient statuses. Episodic responsibility, clearly, is not a resilient status. Since one can be episodically responsible for some things and not others. Your episodic responsibilities will vary with mine, but we are both moral agents. So, being episodically responsible for certain things cannot be a condition of moral agency and the responsibility condition of moral agency must be dispositional responsibility.

What does it take to be dispositionally responsible? It seems nebulous, since different moral agents have different episodic responsibilities. But it can be described more abstractly in terms of episodic responsibilities. Perhaps you are dispositionally responsible when you are episodically responsible for a sufficient proportion of actions, or, failing, that, when you are episodically responsible for a certain degree of consequences (either by being episodically responsible for a few greatly influential actions or many less influential ones). Neither are compelling, but both seem along the right lines.

This rightly alights upon the breadth of responsibility needed for moral agency that is lacking in a condition that refers to episodic conditions. A robot, for example, might reasonably be episodically responsible for cleaning the floor, but they cannot be episodically responsible for very many other things. Humans are generally adaptive and can be episodically responsible for a wide range of things. Thus, they are rightly described as dispositionally responsible.

However, the suggestion that moral agents *must be* episodically responsible is surely false. Like with responsiveness to moral reasons, you may not be *actually* episodically responsible for any actions while retaining a kind of dispositional responsibility. Some people, indeed, never seem to be episodically responsible for anything. But that inability ought not invalidate their resilient status of being dispositionally responsible. Joking aside, consider a prisoner or Prometheus bound to a rock -- they may be unable to act and thus unable to be episodically responsible for anything (they fail the control conditions), but they are still moral agents. They are moral agents, I suggest, because they have *the capacity* to be responsible. That is, they have no episodic responsibility, but they retain dispositional responsibility.

The important thing, if responsibility does bear a conceptual relationship with moral agency, must be dispositional responsibility, and dispositional responsibility is the capacity to be responsible for a sufficient (which I leave undefined, but again we can appeal to human-equivalency if necessary) range of actions. Thus, when some say, ‘one must be responsible to be a moral agent’; they are most charitably interpreted as saying ‘one must be *dispositionally* responsible to be a moral agent’, i.e., ‘one must have the capacity to be responsible for a sufficient range of actions to be a moral agent’.

This idea finds correspondence in wider group agency and moral responsibility literature; Stephanie Collins, for example, says that “‘Moral agency’ implies the capacity to act, bear moral reasons, reason about morality, acquire duties or obligations, and be blameworthy or

praiseworthy." (Collins, 2023). For Collins, being blameworthy or praiseworthy is being responsible, so, she believes that moral agency implies the capacity for responsibility.

I have argued then, that those who Behdadi and Munthe take to claim that 'responsibility is required for moral agency' are best understood as claiming that 'the capacity to be responsible for a sufficiently wide range of actions is required for moral agency'. This claim is not, however, free from opponents. Some claim that moral agency does not require even the capacity to be responsible.

For Christian List, responsibility requires moral reasons-responsiveness (in the form of the capacity for normative judgement) and the further conditions of being well-informed and 'in control' of one's actions (List, 2021). According to List's account, moral agency is nothing more than moral reasons-responsiveness. Thus, for him, moral agency is necessary for the capacity to be responsible. But on this understanding moral agency does not entail that the capacity for responsibility, as one may be entirely incapable of being in control of one's actions or to be well-informed about them even if they can be morally reasons-responsive. So, for List the capacity to be responsible is not necessary for moral agency. Floridi & Sanders' (2004) position is somewhat different. They say that responsibility is a feature of agents with certain intentional states. So, for them, the capacity to be responsible is not necessary for moral agency either, as one can be a moral agent without any intentional states at all, and thus be incapable of being responsible.

The positions are a little closer than the diametric opposites of necessity relations that Behdadi and Munthe identify, but there is a meaningful disagreement here. The issue of contention is whether moral agency involves the capacity for responsibility. A brief attempt at a resolution: it strikes me as unintuitive to separate the capacity for responsibility from moral agency. An agent with adequate moral reasons-responsiveness that does not have the capacity to be responsible seems to me not to be a moral agent at all. For List, at least, this may not be an intolerable idea. On List's account being an agent who is morally reasons-responsive ought to imply two things: first, as an agent, moral agents can *know* things and are thus capable of being 'well-informed'; second, an agent capable of moral reasons-responsiveness is probably capable of meeting the control conditions of responsibility, especially if those control conditions themselves are merely a *general* form of reasons-responsiveness. Fischer & Ravizza (1998), at least, think that moral reasons-responsiveness implies the capacity for general reasons-responsiveness.

Those like Floridi & Sanders who hold fast to the idea that moral agency is necessary but not sufficient for responsibility suggest that it clears up conceptual confusion, I worry, like Coeckelbergh (2014) does, that this comes at the cost of redefining moral agency. Recall 3.2.'s discussion of coherence with intuition and pre-theoretical intuitions. I suggest that the pre-theoretical intuition that moral agency requires the capacity to be responsible is common, and that theories which deny this do not offer sufficient evidence to justify their deviation from intuition. Second, at least for the aim of defending the possibility of artificial moral agents, weaker definitions of moral agency are unlikely to be satisfactory. Adopting an unjustifiably weak definition and then arguing that artificial moral agents are possible will be accused of begging the question. Third, in practical terms the treatment of agents is guided significantly by their capacity for responsibility. If artificial agents are moral agents but unable to be responsible, they will have an eccentric responsibility-free type of moral agency, rather than the familiar human responsible moral agency. This would, pragmatically, make the theory harder to use and perhaps less parsimonious; and that difficulty does not seem obviously counterbalanced by gains in explanatory power.

However, I will not here argue further for this. For my purposes it is sufficient to stipulate that I will focus on 'responsible moral agents' – that is, moral agents that have the capacity to be responsible, regardless of whether that description is implied by the definition of moral agency (or responsibility) alone. The rest of the thesis, then, assumes that to be a moral agent an agent must have adequate moral reasons-responsiveness and the capacity to be responsible.

In summary, the definition of moral agency being used here is the following.

To be a moral agent, an agent must have:

(a) The capacity to be adequately responsive to moral reasons.
(b) The capacity to be responsible for a sufficiently wide range of actions.

Clearly, these conditions do not specify the *degree* to which one needs to be responsive to moral reasons or the range of actions for which one needs to be responsible. Both conditions identify a threshold without explaining *what* that threshold is. Exploring where this threshold truly lies involves a discussion of what kinds of mechanisms fulfil them – and the properties necessary for those mechanisms.

3.6. Why These Conditions of Moral Agency?

Many may think that this definition of moral agency misses out crucial components. Chief among them things like autonomy, free will, consciousness, and intentionality. This is intentional. The definition of moral agency here cuts across these properties and aims to be a guideline and measure for identifying which are necessary for moral agency. The conditions presented here are intended to be a relatively uncontroversial foundational starting point for examining questions about which types of things are necessary for an agent to be a moral agent. So, the absence of potentially controversial necessary conditions of moral agency is an intentional omission.

Of course, I do not expect that everyone will find nothing to disagree with about the two conditions I have proposed. So, I will explain my reasons for doing so a little further. Part of the reason for adopting this approach to moral agency is that artificial systems are *new*. They provide new evidence for revising our pre-theoretical intuitions about moral agency. We should not, then, assume that properties of moral agents are *necessarily* properties of moral agents unless they are *conceptually necessary*. The two conditions I offered are reasonable candidates for being conceptually necessary for moral agency. The properties discussed in the following chapters may be present in all moral agents, or instrumental for moral agency, but they do not seem conceptually necessary for it like the capacity to respond to moral reasons and the capacity to be responsible are.

Chapter 4 and 5 (and to a lesser extent chapters 6 and 7) investigate these properties and attempt to find which are plausibly conditions for being a moral agent. I discuss, in turn, consciousness, intentionality, and autonomy. During this, I argue that artificial systems can be morally reasons-responsive but struggle to do so while also being capable of being responsible. Being capable of responsibility, I argue, requires autonomy, and artificial agents cannot be autonomous because they are designed. Finally, I discuss the evolution of moral agency and suggest a means of developing autonomous and moral reasons-responsive artificial agents: simulating evolutionary processes. These artificial agents, I argue, are the only viable candidate for artificial moral agency.

4. Consciousness Conditions for Moral Agency

4.1. Introduction

Do you need to be phenomenally conscious to be a moral agent? The standard answer, at least in discussions of artificial moral agency, is that you do (Champagne, 2021; Himma, 2009; Moor, 2006; Purves et al., 2015; Sparrow, 2007, 2021; Torrance, 2008, 2014; Véliz, 2021; Wallach et al., 2011). Those that think so support a 'Consciousness Condition'[22] of moral agency. 'Functionalist' accounts of moral agency oppose this and believe that you only need to perform certain functions to be a moral agent, regardless of whether you are conscious or not (Coeckelbergh, 2009; Floridi & Sanders, 2004; Gunkel, 2012; Hakli & Mäkelä, 2019; Ioan & Howard, 2017; Parthemore & Whitby, 2013; Powers, 2013; Søvik, 2022; Tigard, 2021; Tollon, 2021; Wallach & Allen, 2008).

The functionalists' chief argument against the consciousness condition is the 'Epistemic Challenge'[23]. The Epistemic Challenge contends that moral agency cannot have a consciousness condition because humans cannot tell whether other agents are conscious. This chapter defends a modified version of the Epistemic Challenge, and, consequently, rejects the consciousness condition for moral agency.

After introducing consciousness conditions for moral agency in 4.2., I clarify the version of the Epistemic Challenge I focus on in 4.3. and 4.4. Then, 4.5. presents two counterarguments to it. First, the *metaphysical* conditions for moral agency are independent of their *epistemic* status; and second, humans *can* tell whether other agents are conscious. I argue that the first counterargument fails but that the second succeeds. In 4.6., I argue that the Epistemic Challenge's weakness to the second counterargument lies in a failure to acknowledge that humans are specialised to interpret other humans. I suggest that modifying the Epistemic

[22] This has been called the 'consciousness criterion' too.
[23] Behdadi and Munthe (2020) call a similar strategy the 'epistemic argument'; Dung (2022) calls another the 'epistemic objection'. I am not sure if the arguments are identical, so I call this a 'challenge'.

Challenge to focus on the epistemic inaccessibility of non-human consciousness is an improvement on that front. This modification also alters the possible replies to the first counterargument (metaphysics vs. epistemology), but I argue that the first counterargument remains ineffective in this case too.

Then, in 4.7., I present a supporting consideration for the Epistemic Challenge. As argued by David Papineau (2002) and Peter Carruthers (2019), the phenomenal concept strategy of materialism implies that some agents may be neither conscious nor not conscious. Jonathon Birch (2022) argues that if this is true then some animals may neither have nor lack moral status. If so, they would neither fulfil nor fail the Consciousness Condition of moral agency either.

4.2. The Consciousness Condition for Moral Agency

That moral agency has a 'Consciousness Condition' is one part of the so-called 'standard account' of moral agency. Dorna Behdadi and Christian Munthe summarise this: "The standard view of human moral agency is that moral agents must meet rationality, free will or autonomy, and phenomenal consciousness conditions." (Behdadi & Munthe, 2020, p. 197). This section reviews some arguments for the Consciousness Condition.

Kenneth Einar Himma (2009, pp. 23–25) offers two arguments for a Consciousness Condition. First, he argues that only conscious agents can be responsible ('accountable' in his terms) because only conscious agents can be praised or blamed. "[I]t it is conceptually impossible to reward or punish something that is not conscious." (Himma, 2009, p. 25). Second, he argues that only conscious agents can have a capacity 'fairly characterised as moral reasoning' (which I readily interpret as 'adequately morally reasons-responsive') because "the very concept of deliberation presupposes the capacity for conscious reasoning." (Himma, 2009, p. 25) He concludes that moral agency has a Consciousness Condition: "[W]hile consciousness, of course, is not a sufficient condition for moral agency […] it is a necessary condition for being a moral agent." (Himma, 2009, 26).

Himma's arguments for a Consciousness Condition neatly match the two central arguments Behdadi & Munthe identify. They say that defenders of the Consciousness Condition "have presented two main arguments in its favo[u]r:

1. One needs phenomenal consciousness to engage in the sort of decision-making and appraisal that moral agency requires.
2. Phenomenal consciousness is necessary for practices of moral praise and blame to be meaningful." (Behdadi & Munthe, 2020, p. 201)

Further examples of these arguments can be found in papers by Carissa Véliz (2021), Steve Torrance (2008, 2014), and Rob Sparrow (2007, 2021).[24]

Véliz (2021) argues for the Consciousness Condition by arguing that an agent must be conscious to have moral emotions, and they must have moral emotions to adequately respond to moral reasons. She argues for this by appealing to 'philosophical zombies', hypothetical physical and functional duplicates of humans who lack phenomenal consciousness. Véliz argues that philosophical zombies cannot be moral agents. She says, "entities that do not feel cannot value, and beings that do not value cannot act for moral reasons. Moral zombies, therefore, are incoherent. Zombies might act in ways that harm and benefit human beings, but they could never be moral agents or morally responsible." (Véliz, 2021, 8).

Torrance (2014) offers a similar kind of argument. He claims that the "type of rationality that is associated with being a morally responsible, morally reflective agent in the human case, may be seen as being integrally bound up with that human being's sentience." (Torrance, 2014, 510). Rob Sparrow (2007, 2021) argues that artificial agents cannot be morally responsible because they cannot experience conscious first-person emotional states, meaning that they cannot be punished, empathise with others, or truly appreciate moral reasons.

Most arguments for the Consciousness Condition of moral agency, then, take the following form. Humans are moral agents because they can respond to reasons provided by moral emotions, and moral emotions are irreducibly conscious. Thus, moral agency has a Consciousness Condition. This typically leads to suspicion about the possibility of artificial moral agency, since many expect that artificial systems cannot be conscious.

[24] Marc Champagne (2021) advances a similar argument that I will not summarise here, but these examples should suffice to outline the position.

4.3. Epistemic Challenges in Moral Agency and Moral Status

The rest of the chapter discusses a popular argument against the Consciousness Condition. I call this argument the 'Epistemic Challenge'. The Epistemic Challenge argues that there is a considerable epistemic obstacle to knowing whether other agents are conscious and claims that this obstacle means that moral agency should not have the Consciousness Condition.

Before describing this challenge further, I here raise a contextual issue. The Epistemic Challenge about moral agency's Consciousness Condition carries several parallels with arguments about the conditions for moral *status*[25]. Some defend a consciousness condition for moral status (Cochrane, 2018; DeGrazia, 2020; Dung, 2022; Kriegel, 2019; Mosakas, 2021; Sinnott-Armstrong & Conitzer, 2021), and others argue against this condition with a version of the Epistemic Challenge (Coeckelbergh, 2012; Danaher, 2020; Gunkel, 2014; L. S. M. Johnson, 2021; Kagan, 2019; Schwitzgebel & Garza, 2015; Shepherd, 2023; Shevlin, 2021) Such discussions are referred to liberally in this chapter, as the argumentative strategies are very similar. However, the two consciousness conditions are argued for in different ways and are conditions for distinct concepts.

Moral status does not entail moral agency. Children, for example, may not be moral agents, but they certainly have moral status, and the same has been argued for animals (For example, Rowlands, 2012). Moral agency may not entail moral status, either. Some assume that every moral agent has moral status, but others have argued against this (for example, T. M. Powers, 2013).

Defenders of the Consciousness Condition (capitalisation denotes the condition specific to moral agency) use different arguments to defenders of the consciousness condition for moral status. Consciousness is seen as necessary for moral agency because agents need conscious experiences (like moral emotions and/or empathy) to put them into contact with moral reasons and be punishable. The conscious experiences purported to be necessary for an agent to have moral status are different. Sometimes *any* conscious experience is thought to be sufficient for moral status. Both consciousness conditions argue that some conscious experiences are essential, but they focus on different kinds of conscious experience.

[25] An agent has 'moral status', sometimes called 'moral patiency', when it has rights, interests, or is otherwise worthy of moral concern.

While there are clear differences between the two consciousness conditions. The argumentative strategy of the Epistemic Challenge is generally the same. Those who object to consciousness conditions for moral status with the Epistemic Challenge are likely to object to Consciousness Conditions for moral agency on the same grounds. In the following section I canvass several versions of the Epistemic Challenge, including some that are originally intended to target consciousness conditions for moral *status*, and offer the argument.

4.4. The Epistemic Challenge

The first step of the Epistemic Challenge is to claim that humans do not really know about the internal qualities, and particularly the phenomenal consciousness[26], of other minds[27]. The second is claiming that if humans do not have knowledge about other minds' internal qualities when they ascribe moral agency[28], then those internal qualities cannot be conditions for moral agency. Instead, supporters of the Epistemic Challenge suggest that moral agency conditions must be easier to know about, such as an agent's behaviour, functioning, dispositions, and context. Some examples of the Epistemic Challenge follow.

Floridi & Sanders offer what is probably the original formulation of the Epistemic Challenge in the context of artificial moral agency.

> [I]ntentional states are a nice but unnecessary condition for the occurrence of moral agenthood. First, the objection [that intentional states are necessary for moral agency] presupposes the availability of some sort of privileged access (a God's eye perspective from without, or some sort of Cartesian internal intuition from within) to the agent's mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice. (Floridi & Sanders, 2004, 365).

Floridi & Sanders target intentional states, but their argument applies just as well to conscious states. Presumably Floridi & Sanders would agree that 'privileged access' to

[26] There are two related claims that this can represent. First, the claim that we cannot know whether another agent has phenomenally conscious states, and second, the claim that we cannot know the *quality* or subjective feeling of their phenomenally conscious states. It is taken to be the former here – but the latter may yield a more nuanced epistemic challenge of its own, perhaps with similar to 'inverted qualia' thought experiments.

[27] This is occasionally referred to as part of the 'problem of other minds', which seems apt. See (Mosakas, 2021; Tollon, 2021; Torrance, 2014) for references to the problem of other minds in this context.

[28] It is possible to substitute in 'moral status' for the Epistemic Challenge against moral status, per the last section.

conscious states 'cannot be easily guaranteed', and that therefore they are an 'unnecessary condition' for moral agency.[29] Floridi & Sanders' Epistemic Challenge needs one more premise to succeed: states which humans do not have 'privileged access' to must be poor conditions for concepts like moral agency. But Floridi & Sanders assume this, rather than arguing for it.

Discussions about moral status often target consciousness conditions more precisely. In the following passage John Danaher presents a version of the Epistemic Challenge.

> Ought implies can and, apart from the outward behavioural signs, there is no way to confirm or deny the presence of phenomenal states in others. So if phenomenal consciousness is to provide a practicable ground for moral status, it must be because it is cashed out in behavioural terms. It is in this (epistemic) sense that what is going on "on the inside" does not matter from an ethical perspective. (Danaher, 2020)

As per 4.3., although Danaher targets moral status, his Epistemic Challenge also challenges Consciousness Conditions for moral agency. He focuses on the second step: only epistemically accessible conditions can provide 'practicable grounds' for moral status. The first step is left as a conditional, leaving the argument something like the following: if behavioural signs cannot provide sufficient epistemic access to others' consciousness, then consciousness cannot be a practicable ground for moral agency. If, however, behavioural evidence is sufficient evidence to justify the belief that others are conscious, then consciousness may be a condition of moral agency via behavioural evidence. Of course, this latter proposal would be unpopular with standardists, who typically want to emphasise the difference between consciousness and behaviour. It also opens an avenue to reducing, at least in practice, consciousness conditions to behavioural conditions. Nonetheless, defences against the Epistemic Challenge which claim that we can know that others are conscious (based on behavioural evidence in conjunction with other reasons) will be discussed in the following section. Danaher can be interpreted as supporting the second step of the Epistemic Challenge without necessarily endorsing the first.

Coeckelbergh (2009) offers another example of the Epistemic Challenge. Like Floridi & Sanders, he targets intentional states wholesale.

---

[29] Moreover, this interpretation is apt independent of whether conscious states are intentional states, since conscious states are in any case 'mental'.

> The standard account of moral agency and moral responsibility starts from a kind of Cartesian 'mind' problem: we might doubt that our own thoughts, beliefs, desires, etc., are really ours. Is there a demon which deceives me, and pulls the strings? I might falsely believe that I act freely. This doubt is then projected onto 'other minds': do they also act 'freely'? … In real life, however, people seldom contemplate such an issue, and if they did so frequently, they would rightly run the risk of being considered mad by their fellows. Instead, people go on to interact with each other, presuming that the other is a free, moral agent who can and should take moral responsibility. … We do not really penetrate into the 'depth' of their minds, and there is no need to do so in moral practice. We interact with others, treat them, ascribe responsibility to them, and blame them on the basis of how they *appear* to us, not on the basis of what kind of mental states that person really has—if we could even know that at all. (Coeckelbergh, 2009)

Coeckelbergh takes the first step of the Epistemic Challenge by claiming that humans do not 'penetrate into the depth' of others' minds – which I take to mean that humans do not have knowledge of others internal states (like consciousness). But he offers a better developed explanation for the second step of the Epistemic Challenge. Moral agency must have epistemically accessible conditions because it is, Coeckelbergh thinks, *determined* by our ascription practice. Since in practice humans do not know others internal states, but do have an ascription practice for moral agency, those internal states cannot be practicable grounds of moral agency.

The Epistemic Challenge has been used to target consciousness and intentional states. I focus on the Epistemic Challenge that targets consciousness because it is the more parsimonious[30] objection to Consciousness Conditions. Intentional states will be discussed in the following chapter. Here is the specific Epistemic Challenge I focus on.

The Epistemic Challenge for Consciousness Conditions of Moral Agency

1) Humans do not have knowledge about whether other agents are conscious.
2) If humans do not have knowledge about whether other agents are conscious, they cannot make moral agency ascriptions that are sensitive to the consciousness of other agents.

[30] Since an Epistemic Challenge against intentional states needs to explain why all conscious states are intentional states to target Consciousness Conditions; but an Epistemic Challenge against conscious states does not.

3) If moral agency ascriptions are not sensitive to the consciousness of other agents, then moral agency cannot have a Consciousness Condition.

C) Moral Agency does not have a Consciousness Condition

4.5. Objections to The Epistemic Challenge

This section presents two counterarguments to the Epistemic Challenge. In the next section I defend a modified version of the Epistemic Challenge against these counterarguments. Here, I argue that the counterarguments are convincing enough to warrant a response and successfully shift the burden of proof to the supporter of the Epistemic Challenge.

The first counterargument rejects premise 1. Premise 1 claims that humans do not have knowledge *that* others are conscious. Knowledge that others are conscious is independent from knowledge of the *nature* of others conscious states. The counterargument against premise 1 can begin by pointing out that while it seems straightforwardly true that humans do not have knowledge of the nature of other agents' conscious states, it does not likewise seem straightforwardly true that humans do not have knowledge that other agents are conscious. In fact, an objector can claim that humans *do* have knowledge that other agents are conscious – in particular, other humans.

The objector can argue for this by appealing to accounts of how humans navigate the 'problem of other minds' (Avramides, 2020; Hyslop, 2019). For example, one way of justifying the belief that other humans are conscious of others is through analogy. You are conscious and other humans do not seem relevantly different to you, so your belief that they are also conscious seems justified[31]. That there is no relevant difference seems to be backed up by scientific theories, Torrance (2014, p. 20) says, "there is a mass of scientific theory and accreted experimental data linking [consciousness] in humans, and affective properties more generally, with our evolutionary history, and with our current biological and neural make-up." But the belief that other humans are conscious can be justified by more than just analogical inference. That other humans are conscious is reasonably part of the *best explanation* of the world: competing explanations must have it that you are conscious while

[31] In the problem of other minds this is represented by Alec Hyslop and Frank Jackson's (1972) view.

others *are not*, which seems unjustifiably skeptical or otherwise unwieldy[32]. Furthermore, the ability to empathise, perhaps supported by parts of the brain such as 'mirror neurons', may provide a distinct, perception-like[33], means of justifying beliefs about the internal states and consciousness of other humans. Finally, in cognitive science those working on 'Theory of Mind' offer detailed descriptive accounts, such as 'theory-theory' and 'simulation-theory', for how humans come to beliefs about other humans' intentional states (Goldman, 2006; Nichols & Stich, 2003). The same capacities may also justify beliefs that other humans are conscious. There are a range of means to object to premise 1 on the grounds that humans do, in fact, know that other humans are conscious.

One response to this objection is that humans may have good evidence that other humans are conscious, but this does not entail that they *know* it. However, this response fails because it entails an unacceptably high standard of justification for knowledge. Kestutis Masokas argues that adopting this standard "requires us to consider seriously all sorts of bizarre asymmetrical hypotheses—such as, for instance, that one's experiences of the external world and other entities are mere simulations induced by the evil Cartesian Demon… Any such hypotheses are bound to have lower prior probability than the relatively mundane scientific realism". (Masokas, 2020, p. 432). I agree – the response that humans do not *know* that others are conscious is unconvincing. Humans seem to have ample evidence to justify the belief that other humans are conscious.

The second counterargument rejects premise 3 by distinguishing between *epistemic* and *metaphysical* conditions of moral agency. Masokas, for example, claims that the Epistemic Challenge "does nothing to refute the claim that consciousness is morally necessary because it merely concerns our epistemic limitations, and not what constitutes moral status per se." The counterargument is that the truth of premise 1 has no consequences for moral agency's metaphysical conditions. Therefore, consciousness can be metaphysically necessary for moral agency even if humans do not know whether other agents are conscious[34].

Himma presents a version of the objection:

---

[32] Again this mirrors the parallel discussion in the problem of other minds (Melnyk, 1994; Pargetter, 1984)
[33] Similar to Fred Dretske's (1973) account of other minds
[34] Danaher suggests that even if this is true, the ascriptions involved still *in practice* refer to behaviour and so a consciousness condition is not used. Coeckelbergh (2009) can be taken to hold a similar view. Though this is not the response that I will pursue.

> [I]t is not the idea of a moral agency that presupposes that we can determine which beings are conscious and which beings are not; it is rather the ability to reliably determine which beings are moral agents and which beings are not that presupposes that we can reliably determine which beings are conscious and which beings are not. If moral agency presupposes consciousness, then we cannot be justified in characterizing a being as a moral agent unless we are justified in characterizing the being as being conscious. (Himma, 2009, footnote 13).

If we start from moral agency ascription practice, we can draw up a folk concept of moral agency that lacks consciousness conditions. But, Himma argues, moral agency is not determined by ascription practice, it is determined by its necessary metaphysical properties. If our ascriptions do not reflect these metaphysical properties, then it unjustifiably ascribes moral agency. The suggestion is that Coeckelbergh and Danaher are wrong to focus so narrowly on practicable, epistemically accessible conditions because they ought to be focusing on the *real* metaphysical conditions of moral agency.

While discussing consciousness conditions of moral status, Leonard Dung claims that this counterargument refutes the Epistemic Challenge:

> [I]t is perfectly coherent, although it would be unfortunate, that facts about [moral status] are unknowable to us. To illustrate this, consider that water might be said to be grounded in its chemical structure. Even if most objects—including instances of water—could not be tested for their chemical structure, this does not refute the contention that water is determined by H2O. […] Since sentientism [a consciousness condition advocating view] is consistent with the unknowability of sentience in non-human animals and machines, the epistemic objection does not threaten sentientism. (Dung, 2022)

Dung thinks that the metaphysical fact that water is H2O is central to the concept of water. Analogously, the metaphysical fact that consciousness is necessary for moral status is central to the concept of moral status. He claims that whether it is *knowable* that any object is H2O or conscious is irrelevant. The counterargument against premise 3 claims that even if it is unknowable whether any other agent is conscious, moral agency can still have a Consciousness Condition.

Since Dung brought up the H2O example, I turn to a classic case. In Putnam's 'Twin Earth' thought experiment (1975), 'twater' is a functionally water-like substance with the chemical formula XYZ. Suppose, now, the following hypothetical version of Earth, EarthX.

EarthX: On EarthX it is *unknowable* whether any given substance is XYZ or H2O, and half of all water-like substances are XYZ and half are H2O. Furthermore, EarthX's human society

is comparable, technologically, scientifically, and culturally, to Earth's societies at, say, the year 1200. Thus, there is no scientific theory that would draw attention to the unknowability of the atomic make-up of water-like substances.

Humans on EarthX call both XYZ and H2O water and will go on calling both substances water indefinitely. Presumably, Dung would say that it is coherent but unfortunate that half of all water ascriptions are false. But there is something strange about this. Surely the charitable and most coherent interpretation of EarthX's humans is that they use 'water' to refer to both XYZ and H2O. There cannot, by stipulation, be any evidence in their world that H2O is water while XYZ is not. So, they *cannot* refer to H2O alone, at best they can use 'water' to refer to 'H2O or XYZ' or 'water-like substances'. Resultingly, EarthX humans do not falsely use 'water' to refer to XYZ. They correctly use 'water' to refer to 'H2O or XYZ', since all water-like substances on EarthX *are* H2O or XYZ. In fact, what turns out to be false is the metaphysical claim that 'water is determined by H2O'. It seems that the metaphysical conditions must follow the epistemic conditions.

The counterargument against premise 3 is analogous. The metaphysical conditions of moral agency must follow the epistemic conditions that guide moral agency ascription. If it is unknowable that other agents are conscious, then our concept of 'moral agency' cannot refer to consciousness.

An objector might reply in two ways. First, they could deny the unknowability claim and suggest that moral agency refers to consciousness because in some cases we do know that moral agents are conscious. This reply can draw on the earlier counterargument to premise 1. Furthermore, it can perhaps argue that humans have knowledge about one conscious moral agent – themselves – without denying premise 1. However, it is hard to work from your own consciousness to metaphysical moral agency conditions. It is hard to see, from a subjective perspective, why being conscious *makes you* a moral agent, or that if you were *not* conscious you would *not* be a moral agent. Moral agency, after all, is standardly used to describe *others*, not oneself. Drawing on the counterargument to premise 1 seems more convincing.

Second, the objector could argue that there are independent reasons for thinking that moral agency refers to consciousness independent of the knowability claim. As mentioned earlier, supporters of consciousness conditions appeal to thought experiments about philosophical zombies. But these thought experiments beg the question. The intuition that philosophical zombies cannot be moral agents is a consequence of assuming the theory that 'moral agency'

refers to consciousness. That theory, I have argued, can only be only true if other agents' consciousness is knowable, since, as on EarthX, the metaphysical conditions for concepts *follow* from the epistemically accessible conditions. Imagine a scenario in which, by stipulation, one fireplace contains phlogiston, and another does not. A person with a theory of phlogiston has the intuition that there can only be fire in one of these fireplaces, but that is not good evidence that 'fire' refers to phlogiston. The same is true for 'philosophical zombie' thought experiments – the intuition that philosophical zombies are not moral agents is the *result* of the theory that moral agency has a Consciousness Condition[35], not evidence for it. The intuition being tested in this situation is that, assuming consciousness is necessary for moral agency, is a non-conscious agent a moral agent? The answer is clearly no because that is how necessity works. That is not evidence that consciousness is necessary for moral agency.

Attempts to focus on metaphysical conditions alone is bound to fail, since, as seems to follow from 20th century discussions of externalism, metaphysical conceptual conditions are guided by which conditions are epistemically accessible. However, the counterargument to premise 3 may work, but only if it appeals to the objection to premise 1. A consciousness condition can be defended by claiming that humans *do* know that other humans are conscious, and therefore that 'moral agency' can refer to consciousness. This is the counterargument that I will modify the Epistemic Challenge to respond to.

4.6. The Non-Human Epistemic Challenge

The force of both counterarguments to the Epistemic Challenge, I have argued, stem from denying premise 1. Premise 1 claims that humans do not have knowledge that other humans are conscious, which the defender of the Consciousness Condition have good reason to question.

[35] Along with any associated claims about moral emotions or other ways of responding to moral reasons being necessarily conscious.

I think that premise 1 can be narrowed to avoid this. The modification is to change premise 1 to refer to the consciousness of *non-humans*. The change leads to the following 'Non-Human Epistemic Challenge'.

4) Humans lack knowledge about whether non-humans are conscious.
5) If humans lack knowledge about whether non-humans are conscious, they cannot make moral agency ascriptions that are sensitive to the consciousness of non-humans.
6) If moral agency ascriptions cannot be sensitive to the consciousness of non-humans, then moral agency cannot have a Consciousness Condition that applies to non-humans.

C2) Moral agency does not have a Consciousness Condition that applies to non-humans.

The Non-Human Epistemic Challenge holds that there is an epistemic obstacle in humans knowing whether non-humans are conscious. Supporters of the unmodified Epistemic Challenge must support the Non-Human Epistemic Challenge, since the unmodified challenge entails this, narrower, modified challenge. But the Non-Human Epistemic Challenge offers better replies to the counterarguments in 4.5.

In adopting premise 4, the Epistemic Challenge supporter acknowledges the evidence against premise 1. The Non-Human Epistemic Challenge instead targets knowledge about non-humans' consciousness. In 4.5., the counterargument against premise 1 rested on humans' specialised epistemic capacities for justifying beliefs that other *humans* are conscious. Humans do not have equivalent specialised capacities for justifying beliefs that non-human agents are conscious.

Henry Shevlin presents a similar consideration about consciousness conditions for moral status:

> [S]entience-based approaches to moral patiency face serious challenges of both a philosophical and practical character. The most daunting is of course the challenge of how we can ever establish whether a given system is conscious. This is an area of intense debate within comparative psychology, but expert opinions vary wildly; some philosophers and scientist extend consciousness to all vertebrates and some invertebrates, while others have suggested that organisms as relatively cognitive sophisticated as fish may lack the capacity to feel conscious pain. (Shevlin, 2021, p. 464)

Shevlin offers evidence that humans do not, even in theory, agree on whether some non-human animals are conscious. It is even less likely that we possess the required capacities to make ascriptions of moral agency sensitive to the consciousness of non-humans in practice.

The overarching claim here is that even if humans know that other humans are conscious, they are unlikely to know whether non-humans are conscious. However, some disagree, saying that, at least in some hypothetical cases, we have strong justification for believing that a non-human is conscious. Marc Champagne, while discussing artificial responsibility, suggests that we can know that Worf, an alien, is conscious:

> [I]n almost all science fiction narratives, aliens are treated as moral agents. Consider that, while there had to be an episode of Star Trek: The Next Generation devoted to establishing whether Data (an android) is deserving of blame/praise, the idea that Worf (an alien) is deserving of blame/praise literally went without saying. Our mythic projections about extraterrestrials thus reveal strong (and perhaps immutable) intuitions that privilege fleshy tissue." (Champagne, 2021)

If premise 4 is correct that humans do not know whether non-humans are conscious (and for Champagne, consciousness is pivotal for being deserving of blame and praise), then why do they so brazenly assume that Worf is conscious? The answer reveals a little more about humans' specialised capacity to justify belief that other humans are conscious. Worf is a humanoid alien – he is, quite literally speaking, a human with some facial prosthetics. Is it any surprise that he is assumed to be conscious? The likely answer is not that he is made of 'fleshy tissue' – it is that he is literally a human, and benefits from the same epistemic mechanism for justifying beliefs about his consciousness as other humans. On one hand, this can be seen as a failure in acting, the audience is unable to sufficiently suspend disbelief and genuinely attribute alien ontology. If Worf *really* seemed non-human, his consciousness may be under much more scrutiny. Consider some other examples from Star Trek: a 'sentient' gas; animatronic insect-like aliens, or the 'Borg' (a parasitic hivemind). I suggest that is entirely *unclear* to the audience that these aliens, portrayed by non-human figures (or in the case of the 'Borg' not inhabiting a single body at all) are conscious. All told, I think premise 4 is

generally plausible. Despite Worf's apparent consciousness, humans do not, in general, in practice, or in theory,[36] know whether non-humans are conscious.

For those unconvinced, slightly weakening premise 4, for example, to: humans do not know whether *some* non-humans are conscious, would not be catastrophic for the Non-Human Epistemic Challenge. For example, chimpanzees and other advanced mammals like elephants may well trigger the same epistemic mechanism as humans do, meaning that humans can plausibly know whether chimpanzees or elephants are conscious. The important thing here is that this is not true for *all* non-humans – there are some non-humans out there for whom the epistemic mechanism is *not* activated (or is clearly highly unreliable) and humans cannot have knowledge about whether those non-humans are conscious. The Non-human Epistemic Challenge can be targeted at those non-humans specifically without significant losses.

I will discuss another reason to believe premise 4 in the following section but let us assume it is true for now. By accepting premise 4 the Non-Human Epistemic Challenge evades the counterargument to premise 1 because it can be true regardless of whether premise 1 is true. I now turn to defending the Non-Human Epistemic Challenge against other counterarguments.

One counterargument distinct to the Non-Human Epistemic Challenge stems from the weaker conclusion C2. The Non-Human Epistemic Challenge concludes that moral agency cannot have a Consciousness Condition that *applies to non-humans*. The defender of the Consciousness Condition may object that 'moral agency' should only refer to humans – and that moral agency can therefore include a Consciousness Condition even if the Non-Human Epistemic Challenge is true. Call this the 'irrelevancy objection', it will come up shortly.

An objector may reject the Non-Human Epistemic Challenge by appealing to the counterargument to premise 3. As in 4.5., the counterargument to premise 3 can only work if premise 1 is denied, and the Non-Human Epistemic Challenge is compatible with premise 1 being false. So, the counterargument to premise 6 is this: the best theory of moral agency is grounded in the knowledge that other humans are both conscious and moral agents. This theory holds that moral agency has a Consciousness Condition. Humans not knowing

[36] Perhaps Floridi & Sanders' phrase 'possible in theory but not easily guaranteed in practice' is too weak to describe the knowledge in premise 4, but it seems along the right lines.

whether non-humans are conscious is not evidence against this theory. It is only evidence that humans do not know whether non-humans are moral agents.

However, I will argue that this counterargument fails. It fails because it unjustifiably assumes that knowledge about human consciousness determines the metaphysical conditions for moral agency while the unknowability of non-human consciousness does not. This assumption is unjustifiably anthropocentric. I will now explain further.

Prima facie, I take it, non-humans are a *candidate* for moral agency – that is, it is not conceptually necessary that only humans can be moral agents like it is necessary that only humans can be presidents or homo sapiens. Both humans and non-humans, prior to any theory-making or evaluating, are equal candidates for moral agency. Given the equal candidature of humans and non-humans, a theory of moral agency should not contain a condition that is unknowable in non-humans, but knowable in humans.

Here's an analogy. A physicist should not reject quantum mechanics because humans can have more knowledge about Newtonian mechanics. Quantum events are candidates for physical explanations, so the best theory of physics should explain them. A physicist ought not simply conclude that they are unknown under Newtonian mechanics and close the matter. Analogously, supporters of Consciousness Conditions seem in a similar position as a physicist supporting Newtonian physics. Non-humans are candidates for moral agency, so the best theory of moral agency should explain non-humans' moral agency (or lack of it). A philosopher should not simply conclude that non-humans have unknown consciousness and close the matter. The inefficacy of Newtonian physics in explaining quantum particles is reason to disbelieve the theory, and the inefficacy of Consciousness Conditions in explaining non-human moral agency is reason to abandon Consciousness Conditions.

One response to this might be that humans are the paradigmatic moral agents, so theories of moral agency only need to explain human candidates alone. This response converges with the irrelevancy objection by claiming that moral agency is necessarily limited to humans. But here is the problem: why should moral agency privilege humans, rather than 'widening the moral circle' (Singer, 2011) to include the possibility of non-human moral agents? There seems no real principled reason to do so.

Deciding on this involves the following choice. Reform moral agency such that it does not include consciousness, abandoning the Consciousness Condition but adopting a concept that can be accurately ascribed to non-humans. Or continue with a Consciousness Condition that

is only really accurate for humans. The cost of the latter seems to be an exclusionary, anthropocentric concept of moral agency. Several others point out the cost of such exclusionism (Behdadi & Munthe, 2020; Birch, 2017; Danaher, 2020; Gunkel, 2012; Singer, 2011; Torrance, 2014), including that we have caused harm by excluding other humans in the past, and that in the face of uncertainty we ought to adopt a 'precautionary principle' and adopt a liberal attitude to ascription. The counterargument to premise 6 rests on an unjustifiably anthropocentric assumption about moral agency, and, therefore, fails.

The Non-Human Epistemic Challenge offers significant advantages compared to the unmodified Epistemic Challenge. Indeed, I think it presents a convincing case for omitting the Consciousness Condition from moral agency. In the following section, I review a related claim concerning the epistemic obstacles to knowledge of non-humans' consciousness.

4.7. Supporting Premise 4: The Indeterminacy of Animal Consciousness

Further evidence for premise 4 of the Non-Human Epistemic Challenge can be found in metaphysical theories of consciousness. Peter Carruthers and David Papineau have argued that many non-humans are indeterminately conscious. Call this the 'indeterminacy argument'.

The indeterminacy argument claims that there are two equally eligible references for consciousness, and no way to choose between them. Under one referential scheme, only humans (and other higher mammals) are conscious. Under the other referential scheme many animals and perhaps artificial systems are conscious (Papineau, 2002; Carruthers, 2019).

This argument depends on the adoption of the 'phenomenal concept strategy'. The phenomenal concept strategy assumes that materialism[37] is true[38]. It holds that concepts which refer to conscious states (i.e., phenomenal concepts) are a distinct class of concept with few conceptual relations with physical concepts. Despite this, phenomenal concepts are

[37] Materialism here is the theory that the ideal scientific theory consists entirely of physical facts and explanations.
[38] Generally, this is thought to be true because of the causal closure of the physical – there is no opening for any causal influences beyond physical facts.

nonetheless ontologically physical. As Jonathan Birch puts is "we are epistemic dualists without being ontological dualists" (2022).

There are three types of theories about consciousness (following Birch, these terms are from Chalmers, 2018). The 'phenomenal concept strategy' is associated with 'Type-B' materialism. Type-B materialists claim that consciousness cannot be reduced to physical facts. 'Type-A' materialists, in contrast, claim that consciousness can be reduced to physical facts. Finally, anti-materialists claim that consciousness a non-physical substance or property.

Under Carruthers and Papineau's phenomenal concept strategy, phenomenal concepts refer to physical states, but these physical states do not play a role in defining the concept itself. Therefore, while physical states *realise* phenomenal concepts, they are not conceptually related to them; that is, you cannot learn about the physical states that realise perceptual phenomenology by analysing concepts about sight or colour. On this view, discovering whether phenomenal concepts are realised by brains, functional structures, or something else is an empirical matter about which physical states best correlate with our phenomenal concept usage. Papineau and Carruthers claim that our usage of phenomenal concepts yields multiple equally eligible references for consciousness. Since there is no way to choose between these references, they conclude that 'consciousness' is indeterminate. Birch explains it like this:

> We are disposed to apply the concept, in our own case, to states that instantiate both properties. There is nothing in the concept, or in its associated conceptions, or in our use of it, that could fix just one of these properties as the unique referent. They are equally eligible candidates for reference. And yet the distribution of F [high-level functional properties] and N [neuronal correlates] in the natural world may well be very different: N is likely to be specific to mammals for the reasons noted above, whereas F may turn out to be possessed by a very wide range of animals (birds, reptiles, fish, cephalopods, arthropods) which have evolved a different neuronal implementation of the same functional property. (Birch, 2022).

Indeterminacy in this sense means that consciousness *can* be definitively ascribed to humans, who have both F and N. However, some animals have only F, whether they are conscious, then, depends on whether consciousness refers to F or N. But it refers to both F and N equally well, so, at least for now, it is impossible to say whether animals with only F are conscious or not. Birch (2022) argues that if this is correct, then animal moral status, which for him depends on being conscious, is also indeterminate. Likewise, consciousness being

indeterminate in this way would be strong evidence for premise 4 of the Non-Human Epistemic Challenge.

However, appealing to the indeterminacy argument as evidence for premise 4 comes at a cost. The indeterminacy argument assumes that both materialism and the phenomenal concept strategy are true. These assumptions can be objected to. Chalmers (2006, 2018), for example, argues that the phenomenal concept strategy in particular is unlikely to succeed. While Type-A materialists argue that Type-B materialism is too weak and must take physical reduction more seriously (Frankish, 2017; Tye, 2011). I do not aim to settle these matters – it is sufficient to understand the argument here as conditional: *If* one grants the indeterminacy argument, then premise 4 is even more defensible.

That said, I suspect that many supporters of the Consciousness Condition are Type-B materialists who support the phenomenal concept strategy. Under Type-A materialism, consciousness can be straightforwardly reduced to physical states, so the Consciousness Condition would be a physical condition. But arguments for Consciousness Conditions about, for example, philosophical zombies being unable to be moral agents would be unworkable, because under Type-A materialism philosophical zombies are incoherent. Nonetheless, a type-A materialist might deny the indeterminacy argument and argue that moral agency has (physical) consciousness conditions. Torrance (2014) seems to take this view[39]. But Consciousness Condition defending materialists who do not want to take this route will need to face the stronger, indeterminacy argument enhanced, version of premise 4 of the Non-Human Epistemic Challenge[40]. Defenders of the Consciousness Condition may also be anti-materialist, but the position remains a minority view and anti-materialism being true seems an oversized cost for a theory of moral agency.

[39] Which I find quite sympathetic, especially if consciousness is both *functional* and physical. But Consciousness Conditions still seem more trouble than they are worth.

[40] One further option is to support type-B materialism but deny the effectiveness of the phenomenal concept strategy. But exploring that goes too far into metaphysics for my scope.

4.8. Conclusion

The Consciousness Condition for moral agency faces the Epistemic Challenge. I have argued that a version of the Epistemic Challenge: The Non-Human Epistemic Challenge, is more effective than other versions. It argued that, since humans do not know whether non-humans are conscious, then the best theory of moral agency should not include a Consciousness Condition. I defended Non-Human Epistemic Challenge against some counterarguments and think it a convincing reason to avoid Consciousness Conditions in moral agency. The relevant conclusion for artificial moral agency is that artificial agents do not need consciousness to be moral agents.

5. Intentionality Conditions for Moral Agency

5.1. Functionalism, Intentionality, and Moral Agency

This chapter assesses the possibility of an 'intentionality condition' for moral agency. I argue that functionalists can accept that moral agency has an intentionality condition by adopting functionalist account of intentionality. Under this account, artificial systems, and other non-human agents, have intentional states. So, artificial systems can satisfy the intentionality condition of moral agency.

There is a trend in the artificial moral agency literature to refer to 'functionalist' and 'standardist' accounts of moral agency. Standardists emphasise properties like consciousness and intentionality, while functionalists emphasise behavioural and functional equivalence. There are several reasons to avoid this labelling, most prominently that most 'functionalists' are not functionalists about intentional states – which is the standard use of the term 'functionalism'. First on the agenda, then, is disambiguating the sense of functionalism about to be used.

A standard philosophical sense of functionalism is 'intentional state functionalism'. Intentional state functionalists think that to have an intentional state is to possess a state that performs a certain function – that is, a state the initiates some kind of change by responding to inputs with appropriate outputs. The difference between functionalism and behaviourism is that behaviourism claims that the *only* important thing is the demonstrated relationship between the inputs and outputs of the system (which can be thought of as 'stimuli-response' relationships), while functionalism believes the nature of the functional state (the internal state that generates the response to the stimuli) is also important. Call the view that intentional states are best characterised with reference to functional states 'intentional state functionalism'.

But functionalists in machine ethics ('machine ethics functionalists') do not necessarily adopt intentional state functionalism. Machine ethics functionalists think the conditions of moral agency are functional and multiply realisable. This corresponds, for example, with the denial of the consciousness condition of moral agency: so long as the function of moral-reasons

responsiveness is realised, the realising states do not have to be conscious. This is the 'functionalism' contrasted with 'standardism'. There are broadly two varieties of machine ethics functionalism. The first is what John Danaher (2020) calls 'ethical behaviourism', essentially the claim that only behavioural dispositions or other functional equivalence are necessary for responsible moral agency. Wendell Wallach & Colin Allen (2008), Susan Anderson & Michael Anderson (2020), Daniel Tigard (2021) and Ioan Muntean & Don Howard (2017) offer similar positions. The second is 'social relationism', associated with Coeckelbergh and Gunkel (2012; 2012), which appeals to functional equivalence in a social and relational context rather than a purely behavioural or individualistic one.

Machine ethics functionalism does not entail intentional state functionalism. Most machine ethics functionalists think that intentional states cannot be a condition of moral agency because they are not functional states. This is why some types of machine ethics functionalism are called 'mindless morality'. However, a few machine ethics functionalists *also* support intentional state functionalism. Powers (2013), and List (2021) offer accounts of moral agency in which an agent is a moral agent when they have the appropriate functionalist intentional state. This chapter defends Powers' and List's position that moral agents need to have functionalist intentional states to be moral agents; and the associated claim that artificial systems routinely have intentional states.

To summarise, there is a tripartite divide of artificial moral agency theories. First, 'standardists' claim that being a moral agent requires an agent to have certain qualities such as intentionality and consciousness. 'Machine ethics functionalists' claim that being a moral agent only requires an agent to have certain functional states. Among machine ethics functionalists, some support 'intentional state functionalism' and think that functional intentional states are necessary for moral agency. But most machine ethics functionalists think that intentional states are irrelevant to moral agency.

5.2. Machine Ethics Functionalism

Machine ethics functionalists normally reject intentional state conditions for moral agency because of the epistemic challenge. In the previous chapter, I argued that a version of the epistemic challenge that focuses on consciousness was effective. But machine ethics functionalism typically uses the epistemic challenge to target *both* conscious *and* intentional

states. It may seem strange, then, that I consider intentionality and consciousness as separate potential conditions of moral agency. However, as will be shown, there are good reasons to do so. The conditions, in my view, merit different treatments. Moral agency ought not have a consciousness condition but ought to have an intentionality condition.

Let me illustrate the difference in approaches to intentionality conditions between machine ethics functionalism and standard accounts with the example of Data from Star Trek (The Next Generation). Data is a humanoid robotic artificial intelligence. He has a 'positronic' brain that runs on a neural network, allowing him to walk, talk, and act in a way largely indistinguishable from humans. Of course, it is unclear whether Data really has intentional states. After all, no matter how complex Data may be, he is a robot made of wires, positrons, and steel; and that constitution may entail that he cannot have intentional states. If so, he fails to meet standardist intentionality conditions for moral agency. This is despite Data functioning in all respects like a moral agent. Machine ethics functionalism holds the opposite view. For them, Data is a moral agent because he *acts* just like one, and the way he processes stimuli and compares internal representations[41] has the same functional inputs and outputs as responsible moral agents. So, machine ethics functionalism says that Data is a moral agent because of his functional performance.

Social relationism is a type of machine ethics functionalism that takes a different route to the same conclusion. They agree that Data is a moral agent on the grounds of his behaviour and the functional roles his processing mechanisms perform. But this is, on social relationism, not the end of things, because those behavioural dispositions and functional roles are only sufficient for moral agency if they cause Data to be treated as a moral agent by others. In fact, social relationists would claim, the other machine ethics functionalists have things the wrong way round: an agent does not need predetermined and fixed functional roles to be a moral agent, but they only need to hit the somewhat malleable functional performance necessary for being ascribed moral agency by other individuals and society at large. So, in the end, Data is a responsible moral agent because his functioning leads *the crew of the Enterprise* (their spaceship) to hold him responsible and respect his moral agency.

[41]Of course, this looks to the viewer like Brent Spiner making regular human decisions – but is in fact Data making decisions using non-conscious functional states. So, remember to suspend disbelief.

For my purposes, social relationalism can be lumped together with most other machine ethics functionalists. This group is united in sharing two central claims. First, an agent is a moral agent when it satisfies certain behavioural and functional criteria. Second, having intentional states is not necessarily one of those criteria, and so there can be completely 'mindless' moral agents. It is this second claim that I will argue against.

In what sense can it be said that Data does not have intentional states? Powers (2013) distinguishes between 'external' and 'internal' intentionality. 'External' intentionality is the intentionality that a system possesses when the 'intentional stance', that is, the explanatory practice that features intentional terms, produces effective explanations. 'Internal' intentionality is the intentionality a system possesses when they have an internal state that makes intentional explanations effective. Artificial systems have external intentionality, but questionably have internal intentionality.

In this case, saying that Data lacks intentional states can only refer to him lacking internal intentionality. Clearly, intentional terms *can* be used to explain his behaviour, so he has external intentionality. The only plausible claim is that we cannot rightly attribute him internal intentional states. Likewise, since the definition of 'agency' from 3.2 supposes that an agent is a system that can be 'loosely' and 'metaphorically' understood to have intentional states, the same kind of qualification can be applied. To be an agent a system must have *external* intentionality, not *internal* intentionality. Data is, according to the account of agency I am using, an agent regardless of his alleged lack of internal intentional states.

As mentioned, most machine ethics functionalists argue that moral agency need not contain intentional state conditions because of a version of the epistemic challenge (see chapter 4). John Danaher (2020) and Mark Coeckelbergh offer some further explanation.

> To be an ethical behaviourist one does not have to deny the existence of inner mental states, nor deny that those inner mental states provide the ultimate metaphysical ground for our ethical principles. [Because even if there is such a metaphysical property] a sufficient epistemic warrant for believing in the existence of this metaphysical property can be derived from an entity's observable behavioural patterns. (Danaher, 2020).

Danaher wants to hold onto the idea that mental states may form the ground of moral status but claims that behavioural patterns alone provide epistemic warrant for believing that an agent has moral status. Presumably the claim also applies mutatis mutandis to intentionality and moral agency. What type of intentionality does an entity's observable behavioural

patterns offer sufficient epistemic warrant for? The natural and obvious assumption is that it is *external* intentionality. External intentionality is denoted by the explanatory effectiveness of intentional explanations; but internal intentionality is denoted by something *further*. So, Danaher can be interpreted to claim that moral agency requires *external* intentionality but cannot possibly require *internal* intentionality.

Coeckelbergh makes a similar claim in the following passage.

> My suggestion is that we can permit ourselves to remain agnostic about what really goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how AAs [Artificial Agents] appear to us, humans. … Instead of trying to find out the truth about the 'content' of the other's mind, it suffices for our moral practices of moral agency and moral responsibility ascription that we develop our capability to perceive, experience, and imagine the form and performance of the other. (Coeckelbergh, 2009).

Indeed, Coeckelbergh suggests an entire 'quasi' set of intentional states that refer to the 'appearance' of intentional states without *in fact* being those states. In short, he suggests that the only intentional states that matter are those that have *external* intentionality.

Coeckelbergh and Danaher both appear to believe that they can and should explain all the conditions and practice of moral agency without referring to internal intentional states. In the following section I argue that they ought to subscribe to intentional state functionalism, under which there is no need to avoid referring to internal intentional states.

5.3. The Case for Internal Intentionality Conditions

Here, I make the case for internal intentional state conditions in moral agency. Machine ethics functionalists like Coeckelbergh and Danaher (and indeed, Floridi & Sanders) argue for the importance of functional conditions *at the expense* of internal intentional state conditions. But here I argue that including internal intentional state conditions in a functionalist theory of moral agency is advantageous.

List & Pettit and other intentional state functionalists[42] think that internal intentional states are functional[43]. For most intentional state functionalists, Data *has* internal intentional states because he has an internal mechanism that performs the appropriate function. If intentional state functionalists are right, there can be a functional internal intentionality condition in a functionalist theory of moral agency. This is the kind of proposal that List (2021) and Powers (2013) argue for.

Most machine ethics functionalists deny internal intentionality conditions for moral agency because of the epistemic challenge. However, the epistemic challenge is less convincing against internal intentionality conditions than against consciousness conditions, as I will explain. The epistemic challenge's effectiveness here depends on whether humans can access the evidence needed to know that other agents have internal intentionality. As I argued for consciousness in chapter 4, if humans necessarily cannot access this evidence then internal intentionality is not much use as a condition for any practicable concept. If so, there would be no need to refer to 'external' and 'internal' intentional states: external intentionality is the only intentionality there is. If so, moral agency *does* have an intentionality condition; a functionalist and epistemically accessible one. Intentional state functionalists would be perfectly happy with this solution – for them, whether we call intentionality 'external' or 'internal' is less important than whether every intentional state is a functional state, which on this account they are. Reducing internal intentionality is not a means of denying intentionality conditions for moral agency. To meaningfully deny intentionality conditions (or even to be agnostic about them) machine ethics functionalists must leave open the possibility that internal intentionality *is* a condition of moral agency distinct from external intentionality.

In chapter 4 I worked around a similar issue by claiming that the epistemic challenge worked against consciousness (without implying full anti-realism about consciousness) because humans can know that other humans are conscious but cannot access the same kind of evidence about the consciousness of non-humans. But a similar move is less attractive here. Consciousness is identified with reference to *phenomenal experience* – a notoriously hard-to-access quality. It is plausible that humans lack the capacity to access evidence about some

[42] And plausibly interpretivists too.
[43] Jackson and Pettit (1990) talk about 'common sense functionalism', which I believe is essentially the same idea; List & Pettit support this approach in their book and describe themselves as functionalist. This is also the position of varied and influential philosophers such as Daniel Dennett, Donald Davidson, and Jerry Fodor.

agent's consciousness because that evidence is unusual: it is about phenomenal properties, which (at least for some) are non-physical or otherwise deeply mysterious. Theories of intentionality, on the other hand, typically hold that internal intentional states are determined by an agent's *history*, *constitution,* or *structure*[44] – all relatively non-mysterious and epistemically accessible physical and empirical qualities[45]. In chapter 4 I suggested that humans have specialised epistemic equipment to know that other humans are conscious and lack that equipment for non-humans. But the same suggestion for internal intentionality is harder to defend because under most theories of intentionality humans *do* have access to the evidence necessary to accurately ascribe internal intentionality to every kind of agent.

The epistemic challenge against internal intentionality conditions is weaker than the one against consciousness conditions because it cannot appeal to genuine epistemic limitations. In practice, humans may be reluctant to ascribe internal intentionality to non-humans. But this reluctance is not, it seems, grounded in a lack of *evidence*, it can only be grounded in different beliefs about what constitutes internal intentionality and (occasionally) anthropocentric bias. If this is true and humans can access all the relevant evidence about internal intentionality, then the epistemic challenge does not get off the ground. This is doubly so if intentional state functionalism is true – as a functional concept internal intentionality would be (almost necessarily) epistemically accessible.

If this is right, then the idea that observable patterns only gave epistemic warrant for beliefs about *external* intentionality turns out to be mistaken after all. Observable patterns, assuming this includes the empirical evidence available in general, *ought*, on most theories of internal intentionality, to provide that epistemic warrant for internal intentionality.

So, the epistemic challenge is probably not a good reason to deny internal intentionality conditions. What other motivations might the machine ethics functionalist have to deny them? One idea is that internal intentionality is simply irrelevant to moral agency ascriptions.

---

[44] I refer generally to 'most theories of intentionality', drawing on general categorisations of such theories such as in Haugeland's (1990) landmark paper 'the all-stars of intentionality' and Hutto & Satne's (2015) "the natural origins of content". The theories of intentionality discussed in those papers are naturalist and, according to Hutto and Satne, interpretivism and teleosemantics are currently the most popular – both of which are, incidentally, plausibly versions of intentional state functionalism (as I will discuss further in the next section).

[45] I assume that there is nothing that 'it is like' to have internal intentionality. That assumption is standard but not universal. Theorists of phenomenal intentionality ought to be concerned about an Epistemic Challenge against intentionality.

But there is much further explanation needed here – as this surely depends on the theory of intentionality you favour. Under an intentional state functionalist account, whether an agent has internal intentionality is highly relevant to that agent's ability to be responsible and respond to moral reasons – without the right internal functional mechanism, the agent may well fail to be adequately moral-reasons responsive. Under some versions of what Haugeland (1990) calls 'neo-cartesian' or 'neo-pragmatic'[46] theories of intentionality, internal intentionality *may* be irrelevant to moral agency. So, this is probably the crux of the machine ethics functionalist position: until a theory of intentionality is proven, it may be better to sit out the debate on whether there can be internal intentionality conditions for moral agency.

So, the machine ethics functionalist has washed their hands of the matter. This is certainly not a bad idea, but they should be prepared for various outcomes: *what if* internal intentionality turned out to be both identifiable through epistemically accessible evidence and non-functional? Let us suppose a mind-brain type identity theory is proven to best fit all the empirical evidence: having an internal intentional state turns out to be having an internal state with a particular structure unique to biological systems. Can the machine ethics functionalist sit back and keep their hands clean? I wonder if they can. Because it seems to me that the mind-brain type identity theorist (now vindicated) is going to be tempted to say that 'as-if' external intentional states are not enough for moral agency because an agent can only *truly* respond to moral reasons if they have internal intentionality. With the empirical evidence on their side and the conceptual ground ceded earlier on, it would be hard to deny them. The machine ethics functionalist might, in this scenario, look back and wonder where they went wrong. Adopting intentional state functionalism *now* gives machine ethics functionalists the tools to disagree with the mind-brain type identity theorists on a conceptual level. They can develop and defend an understanding of intentional states whereby internal intentional states *cannot* be limited to brains or brain-like structures. This is a prophylactic move, and therefore perhaps not a hugely appealing justification for what is a reasonably sized pivot for some machine ethics functionalists. But I think it is a good reason (especially while the option remains open) to adopt intentional state functionalism – at least for those who feel attracted to a functionalist theory of moral agency.

---

[46] Though as Hutto and Satne (2015, p. 528) say that "many regard neo-pragmatism's … strategy as hopeless", I do not consider it a mainstream theory of intentionality.

There are further pragmatic reasons for the machine ethics functionalist to adopt intentional state functionalism. It's a popular and well-defended theory of intentionality; it allows talk of intentional states in real terms, avoiding the 'quasi', 'as-if', and 'metaphorical' qualifications, which complicate things; and it results in a naturally coherent position: if both machine ethics functionalism and intentional state functionalism are true, then moral agency can have a straightforward, functionalist, internal intentionality condition.

I have offered some reasons to think that the neatest move for machine ethics functionalism is to include a functionalist internal intentionality condition. What remains to be seen is whether artificial systems can satisfy that condition. Fortunately, accounts of group agency have outlined just how intentional state functionalism may apply to groups, and the same line of reasoning can be adapted to apply to artificial agents. In the following section I present this account.

5.4. Collective and Artificial Agents' Intentional States

Machine ethics functionalism that adopts intentional state functionalism can embrace an internal intentionality condition for moral agency. On this view, agents have internal intentional states when they have functional states that are usefully interpreted as intentional states. The following sections discuss the extent to which artificial systems have internal intentional states. The discussion will be guided by an analogy between artificial and collective agents. Note that from here on I drop the 'internal' qualifier for intentional states, though it still applies.

In this context, collective agents (which I will also call 'groups' or 'group agents') are a natural parallel to artificial systems, both are non-human agents with the potential to be embedded in our social and political worlds. So, it is prima facie reasonable to suppose that they will be subject to intentional states in the same way. List suggests this parallel and that artificial systems have intentional states.

> In brief, the parallel lies in the fact that group agency and artificial intelligence each involve entities distinct from individual human beings that qualify as intentional agents, capable of acting more or less autonomously in pursuit of certain goals and making a difference to the social world. (List, 2021)

Furthermore, the most popular group agency account (influentially advanced by List & Pettit, 2011) uses intentional state functionalism to argue that groups have moral agency. Some

suggest the same attribution can be analogously made to artificial systems (Collins, 2023; Laukyte, 2017; List, 2021) This section explains how intentional state functionalism has been applied to groups, and how it might analogously apply to artificial systems.

List and Pettit (2011) use intentional state functionalism to attribute intentional states to groups. List claims that intentional state functionalism can be supported by interpretivist evidence and that it supports the claim that artificial systems have intentional states.

> [W]hile from an interpretivist perspective successful interpretability as an agent is constitutive of agency, for me it is merely indicative. That an entity is interpretable as an agent is good evidence for the hypothesis that it satisfies the agency conditions. Irrespective of whether we go with an interpretivist criterion or my preferred realist one, however, the fact that sophisticated AI systems and suitably organized collectives meet [List's functionalist agency conditions] supports the claim that they each qualify as agents. (List, 2021)

Kendy Hess goes further, suggesting that all standard accounts of intentional states can support groups having intentional states: "[C]orporate commitments—however they arise—are easily and casually spoken of as beliefs and desires: … I suggest that these commitments literally qualify as beliefs and desires on the standard interpretationist, dispositionalist, and representationalist accounts developed to explain human belief and desire." (Hess, 2014)

Analogously attributing intentional states to artificial systems will face some challenges. While there are evidently several similarities between collective and artificial agents, there may be significant dissimilarities that undermine the attribution of intentional states to artificial systems. One that can be set aside immediately is consciousness. I have already rejected the inclusion of consciousness in an account of moral agency in chapter 4. Furthermore, there does not seem to be anything central to intentional states that requires consciousness. Jackson and Pettit (1990, 37) have it that "beliefs and desires do not have qualia -- or if they do, they are not essential to the states being beliefs and desires.". A claim reproduced by Kendy Hess: "The standard, widely accepted accounts of belief or desire simply do not include a phenomenal aspect" (Hess, 2014).

So, can it be claimed that artificial systems have intentional states under both interpretivism and representationalism[47]? I will argue for a positive answer.

Interpretivism (despite some misinterpretations that it necessarily implies the 'loose' and 'metaphorical' instrumentalism I described in 3.2.) ascribes *internal* intentional states to agents that behave in a way that can be effectively explained using intentional terms. A focal point of the position is Daniel Dennett's 'intentional stance'.

> You explain to a companion what your laptop computer is doing by noting that it *wants* to print a document, *discovers* that the printer is out of paper, and *hopes* to attract your attention with a bouncing icon. The italicized words in the preceding sentence are not, if Dennett is right, used metaphorically: they apply literally to your laptop computer; you are using them in exactly the sense you would use them in describing the behavior of a fellow human being. (Heil, 2019, p. 156)

Interpretivism is not, strictly speaking, a version of intentional state functionalism, but it occupies a similar space. As List suggested earlier, if an agent has a functional state that appropriately corresponds with the intentional state, then intentional explanations are guaranteed to be sufficiently explanatorily effective. Generally, intentional states under intentional state functionalism are also intentional states under interpretivism[48]. Adopting interpretivism without intentional state functionalism makes little difference to the machine ethics functionalist position – except that it is even easier to justify intentional state attributions to performatively equivalent artificial agents.

Artificial systems, according to most interpretivists, do have intentional states. Raul Hakli and Pekka Mäkelä say: "[I]t seems that most roboticists and a large number of philosophers, some of them inspired by Daniel Dennett's idea of intentional stance, are already now willing to grant at least first-order intentional states, basically beliefs, desires, and intentions, to robots" (Hakli & Mäkelä, 2019). And under interpretivism these are not 'quasi' or 'metaphorical' intentional states. Interpretivism is deliberately formulated to ascribe intentional states to all kinds of systems, including artificial systems, so it is essentially a free win for artificial agents.

---

[47] Unlike Hess, I will not discuss dispositionalism here, because I take it that if artificial systems can have representationalist and interpretivist mental states, then they can also have dispositional ones.

[48] There is nuance here, in that sometimes one can be effectively interpreted without having a corresponding functional state or that one can have an intentional-state corresponding functional state without being effectively interpreted; but these are exceptional cases.

Representationalist accounts of intentional states[49] offer only slightly more opposition to the ascription of intentional states to artificial systems. Here, too, there is a good case for holding that artificial systems have intentional states. Just as Hess says that groups do:

> Representationalists argue that belief requires both an internal representation and (essentially) a disposition to act on the basis of that representation. Dretske (1988) describes such internal representations as "maps by which we steer" (79); a belief is thus an internal representation of the world (or some aspect of it) that guides our behavior. Given how easily corporate agents meet the "disposition to act" requirements, the main question left is whether they can have "representations" in the appropriate sense. So what is it to have a representation? Taking Dretske's (1988) account as our exemplar, a representation is (1) an information-bearing state (2) internal to a system (3) that the system synthesizes from information gathered from the world. Dretske uses the example of a wolf tracking a crippled caribou: The wolf has the ability (via its senses) to take in information about the external world and "indicate" that information to a central system; in this case, information about visual, aural, and olfactory matters. The central system synthesizes this disparate information into a representation that both signifies "crippled caribou" and identifies its relevance to the wolf. Thus the complete representation is more likely "easy prey" or just "dinner." … From this account it seems unproblematic to say that corporate agents have representations. (Hess, 2014)

Powers (2013) also argues that under the computationalist theory of mind, artificial systems possess intentional states by drawing on Dretske's account. Indeed, artificial systems can satisfy the conditions for representationalist mental states more easily than groups do. There are fewer questions about where the central decision-making processes are located and whether those processes are sufficiently interdependent to be a coherent intentional state. An artificial systems' representation is in the circuits, server, positron brain, or other mechanism of the artificial system's memory and decision-making processes. Artificial systems operate on representations that are sometimes explicitly represented (and can be accessed by others) and sometimes merely implicitly represented (which as Hess mentions, is entirely sufficient for the purposes of 'representation') in the functioning, reasons responsiveness, and actions of the artificial system.

One theory of intentionality absent in collective agency literature is teleosemantics. Again, the intentional states are held to be functional, but their function is thought to be fixed by

[49] This type of account is sometimes called functionalism. For obvious reasons I do not conform.

their evolution (Millikan, 1984; Neander, 2017). Here, it is more challenging to say that artificial systems have intentional states, because they have no evolutionary history. One might supplant their design-history, causing their function to be *whatever they were designed to do*. This is something I discuss a little further in chapter 7. If a teleosemantic account of intentionality is true[50], then it is unlikely that artificial systems have intentional states.

So, if we are representationalists or intepretivists it seems that artificial systems routinely have intentional states. If intentional state functionalism is true, then the 'agnosticism' about 'inner mental states' in previous sections can be avoided. As under intentional state functionalism it is possible to know, through their behaviour and functioning, what another agent's intentional states are. It is thus possible to understand machine ethics functionalism as involving, referring to, and requiring, intentional states. Such a position is the standard in group agency.

5.5. Collective and Artificial Moral Agency

So, what is the positive picture of moral agency here? It is that artificial systems have intentional states, and that to be a moral agent one needs appropriate intentional states that represent moral reasons (and are appropriately motivating). Thus, on this account, artificial systems fulfil the adequate moral reasons-responsive condition of moral agency if they have (enough) intentional states that represent moral reasons.

List (2021) presents similar arguments for this use of the analogy between artificial and group agents to attribute moral agency to artificial agents. Though he presents this solely as a positive proposal, rather than one that competes with the standard and functionalist accounts. As a positive proposal, I suggest that it encounters the fewest problems.

The analogy has also been made by Migle Laukyte (2014, 2017), who argues that artificial systems are moral agents according to the criteria of group moral agency. However, Laukyte claims that artificial agents and groups have 'quasi' intentional states, not intentional states

[50] And is the *only* account of intentionality that is true – i.e., there are no hybrid accounts, for example, it is appealing to think that teleosemantics explains the *generation* of intentional states in nature without accurately explaining *the conditions* under which a system has intentionality – a proposal that can be traced at least to Haugeland (1990).

proper. Laukyte attributes this view to List and Pettit, though List and Pettit (based on Pettit's views on the issue (Jackson & Pettit, 1990) and List's (2021) earlier passage) seem inclined to ascribe internal intentional states to both types of agent.

There is a disanalogy between groups and artificial systems: groups can respond to moral reasons through being composed of human members while artificial systems cannot. Groups make moral decisions using moral reasons that are available to the group through human capacities for moral reasons-responsiveness. This does not entail that the group's responsibility or mental states can be reduced to the group members responsibility or mental states, but it does entail that the group's moral epistemology (or 'moral psychology') necessarily involves the individual group member's moral reasons-responsiveness.

For example, suppose there is a workers' union: United Employees (UE). UE protects the rights of workers who are not members of UE. It does so based on the moral reason that "it would be morally abhorrent to protect the interests of union members when there are no relevant moral differences between union members and non-union members". Whether or not this is a sound moral reason is irrelevant, as the question here is: how does UE come to be independently[51] sensitive to this reason? The answer, I suggest, must be that the moral reasons-responsiveness of individual members are a necessary for UE's independent sensitivity to the reason. There must be one or several members with intuitive or emotional reactions that put them into contact with the moral reason above, and which they transmit through UE's epistemic mechanism (say, by reasoned democratic debate in which the moral reason is discussed and accepted by other group members). UE, in this case, is sensitive to the moral reason *because* it contains humans with moral reasons-responsiveness. In contrast, a union of psychopaths would not be able to respond to moral reasons in the same way, because no members would have the required moral reasons-responsiveness (similarly a group whose decision-making mechanism filters out moral reasons could not be moral reasons-responsive). Groups that are moral agents can respond to moral reasons because their individual members are sensitive to those reasons.

[51] It can be 'dependently' sensitive to moral reasons by taking on external moral advice or deferring to moral rules or laws – but moral agency requires independent sensitivity, as will be discussed further in the following chapter

So, here is the disanalogy: artificial agents cannot have moral-reason responsiveness capacities as part of their moral decision-making process like groups can. Artificial systems are like (a group of) psychopaths in that regard. Artificial systems appear to be unable to have their own capacity to respond to moral reasons, and thus, we might think, are unable to be moral agents.

One might wonder *why* artificial agents cannot have these capacities for moral reasons-responsiveness in their own right. After all, according to intentional state functionalism artificial agents have internal intentional states. So why can they not also have moral reasons-responsiveness capacities like humans can? The answer is, at least from my perspective, that there is nothing in principle preventing them from at least having equivalent moral capacities (As argued for by Björnsson & Hess, 2017). Moral reasons-responsiveness is a functional state, just like any other intentional state, so there is no missing quality of 'consciousness' or 'intentionality' behind artificial systems' lack of adequate moral reasons-responsiveness. However, it is also seemingly true that contemporary artificial systems do not yet have them.

Moral reasons are an altogether more nuanced and challenging type of reason to respond to compared to ordinary reasons. While it makes sense to interpret a contemporary artificial system as having a belief or desire, it often will not make sense to interpret them having a *moral* belief or desire. Contemporary artificial systems do not act in a way that reflects moral reasons as consistently as humans (or perhaps groups) do. They may be able to act morally in some limited circumstances, but they do not normally show sufficiently sophisticated behaviour to be considered adequately responsive to moral reasons.

It may be possible (but not yet achieved) for artificial agents to have adequate moral reasons-responsiveness (as will be discussed further in the following section). But, even so, artificial agents cannot respond to moral reasons through the *same* means as a group agent can, because they cannot be constituted by groups of humans[52] in the same way.

[52] Some (e.g. List, 2021) see collective agents as a type of artificial agent, if so, then I exclude them from my discussion here.

5.6. Conclusion

The disanalogy between artificial systems and groups moral reasons-responsiveness is why, in a recent paper, Collins (2023) claims that artificial systems cannot be responsible, while groups can. Humans, and groups of humans, can respond to moral reasons while artificial systems do not, it seems, have equivalent moral competency.

The disanalogy between groups and artificial systems implies that artificial systems are missing a certain functional state. Group agents, such as UE, can have functional equivalents of human moral reasons-responsiveness *through* being constituted by human agents; artificial agents cannot develop moral reasons-responsiveness in the same way. Instead, to be moral agents, artificial systems need to respond to moral reasons in other ways. The traditional suggestions for doing this are to use moral rules as the basis for their moral capacities, or to generate their own moral capacities by inferring the appropriately moral reasons-responsive action from the actions of others, or indeed, from empirical evidence about the world in general (Allen et al., 2005). No artificial system has yet done this sufficiently to be considered to be adequately moral reasons-responsive. But such a thing seems quite feasible and will be explored further in the following chapter.

Part II: Troubles with Moral Reasons-Responsiveness & Autonomy

6. The Moral Decision Machine and Moral Deference

6.1. Introduction

Moral agents must be adequately responsive to moral reasons and responsible. Some argue that artificial systems can never be adequately responsive to moral reasons. Two popular arguments for this were rejected in part I. If moral agency does not have a Consciousness Condition, then the argument that artificial agents cannot adequately respond to moral reasons because they cannot be conscious fails. If artificial agents can and routinely do have internal intentional states, then the argument that artificial agents cannot adequately respond to moral reasons because they cannot meet intentionality conditions fails.

However, there are two more arguments that artificial systems can never be adequately responsive to moral reasons. First, Duncan Purves, Ryan Jenkins & Bradley J. Strawser (2015) argue that if morality is uncodifiable, then artificial systems cannot adequately respond to moral reasons. Second, various other 'standardists' have claimed that artificial systems cannot be adequately responsive to moral reasons because they lack certain metaphysical qualities, such as evolutionary history, emotions, or human-like ontology.

I present a counterexample: The Moral Decision Machine. I argue that the Moral Decision Machine can adequately respond to moral reasons. I then defend this, and the possibility of the Moral Decision Machine, against several objections. If the Moral Decision Machine is possible and adequately respond to moral reasons, then the two arguments above fail.

Finally, I argue that the Moral Decision Machine is not a moral agent for a different reason. It is not a moral agent because it cannot be responsible. It cannot be responsible, I argue, because it is dependent on moral deference. I generalise the argument to claim that many artificial systems depend on moral deference and therefore cannot be responsible.

6.2. The Moral Decision Machine

Even if we accept that artificial agents can meet intentionality conditions and do not need to meet consciousness conditions to be moral agents, there is still plenty of resistance to the idea that they can be adequately responsive to moral reasons.

The first argument for this is 'the anti-codifiability argument'. The anti-codifiability argument is based on the anti-codifiability thesis. "The codifiability thesis is the claim that the true moral theory could be captured in universal rules that the morally uneducated person could competently apply in any situation. The anti-codifiability thesis is simply the denial of this claim, which entails that some moral judgment on the part of the agent is necessary." (Purves et al. 2015) According to Purves et al. the anti-codifiability thesis is a common view entailed by many moral theories, especially moral particularism. Though they admit that if the anti-codifiability thesis is false then the anti-codifiability argument will fail.

For the sake of argument, let us assume that the anti-codifiability thesis is true: the true moral theory cannot be captured in universal, easy-to-apply rules. Consequently, let us accept that an agent must possess moral judgement (or some equivalently uncodifiability-busting capacity, as I will later discuss) to adequately respond to moral reasons. Artificial agents, Purves et al. claim, cannot make moral judgements. They outline several possible understandings of moral judgement and argue that all are barred from artificial agents, which must, as Hubert Dreyfus (1992, p. 199) argues, be "either arbitrary or strictly rulelike". So, without the ability to make moral judgements, artificial agents find themselves unable to adequately respond to moral reasons.

The second argument comes in various forms. The general idea is that artificial agents lack some metaphysical quality necessary for adequate moral reasons-responsiveness. I have already discussed consciousness, but consciousness is a commonly used property in this type of argument. Other versions suggest that adequate moral reasons-responsiveness requires biological functioning, evolutionary history, or emotions. This chapter presents a counterexample of an artificial agent that *can* adequately respond to moral reasons, therefore, whatever metaphysical quality is used in this argument, the counterexample shows that it is either unnecessary for adequate moral reasons-responsiveness, or possible for artificial agents to possess.

Here is the counterexample: 'The Moral Decision Machine'. According to all the arguments just mentioned, The Moral Decision Machine, as an artificial agent, should be unable to adequately respond to moral reasons.

The Moral Decision Machine: In the distant future, a team of computer scientists aim to create a machine that responds to moral reasons. The machine's designers collate a database of human moral decisions. This database, drawn from hundreds of years of human history, contains a series of 'snapshots' of human moral decisions, and there is a snapshot for almost any circumstance. The snapshots consist of empirical information about the relevant (and often irrelevant) physical facts surrounding the moral decision, including the context, relationships, and role of the actor. Then they program a computational artificial system, the Moral Decision Machine, to do the following. First, it gathers the empirical information from a live situation in the same way the snapshots were gathered, with an array of sensors, internet searches and memory banks. Then, the Moral Decision Machine trawls the snapshot database to find a relevantly identical empirical situation. It then acts in exactly the way the human in its historical database acted.

Later, I will argue that the Moral Decision Machine is not a moral agent. However, it does seem adequately moral reasons-responsive. Furthermore, if it is adequately moral reasons-responsive, then it is despite the anti-codifiability thesis being taken to be true. The Moral Decision Machine overcomes codifiability concerns by directly reproducing human moral decisions without codifying any moral principles. It responds to moral reasons identically to humans, so, since humans are moral agents, it adequately responds to moral reasons. If so, it does this despite lacking human ontology, evolutionary history, or emotions (or consciousness). The next section considers two sets of objections: objections that the Moral Decision Machine does *not* adequately respond to moral reasons, and objections that the Moral Decision Machine is impossible.

6.3. Objections to The Moral Decision Machine

In this section, I predict a few objections against the claim that the Moral Decision Machine can adequately respond to moral reasons and offer replies to them.

Objection 1. The Moral Decision Machine does not respond to the *same moral reasons* the human responded to.

There is a good case to be made for The Moral Decision Machine responding to the same moral reasons that the historical human did. If 'ethical supervenience' – which is well-defended in metaethics (see McPherson, 2022), is true, then moral reasons are grounded in a situation's empirical facts. Those empirical facts are what The Moral Decision Machine responds to. In the historical snapshot, a human identified the moral reasons that arose in the empirical situation and responded to them. The Moral Decision Machine responds to a relevantly identical empirical situation with actions that respond to the same moral reasons. So, the Moral Decision Machine acts for the *same* reasons that the human did, even though it might not be able to express, identify or understand those reasons.

Suppose, for example, the Moral Decision Machine is in a hostage negotiation situation. There are many complex moral reasons involved, reasons about the welfare of the hostages, besiegers, and the hostage-taker themselves; reasons about the collateral damage and responsibility attribution; reasons about the morally appropriate solutions and future moral consequences, etc. The Moral Decision Machine analyses the full *empirical* situation: there is one hostage with these characteristics, the hostage taker looks like this and has that history; there are three entrances to the location; it is 2am; there is media interest in the case; etc. Then, it trawls through the database and finds an empirical situation that matches *all* these characteristics – this is the 'snapshot'. To explain what it means to be 'relevantly identical': if you travelled between these cases, it would be like travelling between close parallel worlds – the people would be different, but they would look similar and have similar competencies and histories, the location might be different, but the relevant features of the location would be the same. In the snapshot, there was a human hostage negotiator (who may or may not have been exceptionally competent) who responded to the complex situation, including the moral reasons involved, by performing a series of actions. This series of actions is what the Moral Decision Machine performs. After significant changes in the empirical situation, the Moral Decision Machine searches for snapshots that best match the present state of affairs, and repeats the process. In practice, this would look like the Moral Decision Machine behaving, including explaining its actions and instructing others, like a human hostage negotiator. Its actions respond to all the moral reasons in the situation just as well as the human hostage negotiator did.

Objection 2. The Moral Decision Machine responds inconsistently to moral reasons.

It is true that The Moral Decision Machine will have variance in its moral reasons-responsiveness. It may act inconsistently over time by first responding strongly to one type of moral reason, then responding strongly to a different type.

Imagine the hostage negotiation again; suppose that the initial empirical situation involved a discussion with the hostage taker over megaphone – in the snapshot, the hostage taker surrendered in the face of careful and exceedingly cautious tactics from the hostage negotiator. But, faced with the same tactics from the Moral Decision Machine, the present hostage taker refuses to surrender. At this point, the Moral Decision Machine searches a new snapshot, because the empirical situation has diverged from its initial one. Suppose the Moral Decision Machine's new snapshot centres on a particularly aggressive and gung-ho hostage negotiator who immediately plans to storm the building.

This behaviour seems unusual for an adequately morally reasons-responsive agent, since humans, for example, tend to consistently emphasise the *same* moral reasons over time. Perhaps they are cautious and value minimising average expected harm, or perhaps they are aggressive and value limiting the maximum possible harm by ending things quickly. But they do not tend to be cautious one minute and aggressive the next. The Moral Decision Machine, however, might be.

However, this does not mean that the Moral Decision Machine is inadequately morally reasons-responsive. In most cases, it will respond to the moral reasons in the situation competently and appropriately. There is unlikely to be massive variance in different humans' responses in empirically similar snapshots. Most hostage negotiators will respond to roughly similar moral reasons, and the Moral Decision Machine may perhaps seem erratic, but it would not seem morally incompetent or inadequately moral reasons-responsive.

Objection 3. The Moral Decision Machine would occasionally be incompetent.

Some of the Moral Decision Machine's actions might be incompetent if the human its snapshot centres on acted incompetently. They might commit a big mistake, and the Moral Decision Machine would repeat that mistake.

However, occasional incompetence is consistent with the Moral Decision Machine's overall capacity to adequately respond to moral reasons. Occasional incompetence does not undermine an humans' adequate moral reasons-responsiveness or moral agency. If the Moral Decision Machine is on average incompetent, it is because the average human is less morally

reasons-responsive than expected. However, the average human, by definition (see 3.4.) adequately responds to moral reasons, so the Moral Decision Machine adequately responds to moral reasons.

Objection 4. The Moral Decision Machine is not using moral judgment; therefore, it cannot respond to moral reasons.

The Moral Decision Machine reproduces a historical moral judgement. It acts for the same reasons as the historical human did, and therefore simulates the process of moral judgement. The initial moral judgement was performed by the human in the snapshot, but a considerable amount of the framing, processing, and acting involved is performed by The Moral Decision Machine. Together, these two processes enable the Moral Decision Machine to adequately respond to the moral reasons present in it's situation. One of those processes is a moral judgement, so, it is incorrect to say that the Moral Decision Machine does not *use* moral judgements – it just does not use moral judgements *alone*, which seems unproblematic.

Even if the Moral Decision Machine does not use moral judgements, per se, its means of responding to moral reasons ought to alleviate anti-codifiability worries. It does not oversimplify moral reasons with arbitrary rules or random decisions. The anti-codifiability worry is that rules and random decisions are insufficient to adequately respond to moral reasons, it need not entail the stronger claim that the *only* way of adequately responding to moral reasons is through moral judgement. If the moral decision machine does not use moral judgement, then it seems to me that moral judgement is unnecessary for adequate moral reasons-responsiveness.

Objection 5. The Moral Decision Machine is not using its *own* moral judgement.

There seems to be nothing wrong in principle with using tools to respond to reasons. If I use a calculator to respond to a reason, my actions are still *my own*. The Moral Decision Machine uses human moral judgements in the same way as a human uses a calculator – it outsources some of mechanism for moral reasons-responsiveness. Adequately responding to moral reasons need not require an agent perform *all* that process. My responding to a moral wrong that I view on television, for example, does not undermine my moral reasons-responsiveness despite my reliance on technology.

The motivating concern behind this objection is perhaps that the Moral Decision Machine would fail to adequately respond to moral reasons if it were cut off from its database. This is

true. But the database is *internal* and a central part of the Moral Decision Machine's functioning, it is 'its own' as much as humans' brains are their own. The suggestion that humans are not adequately morally reasons-responsive because lobotomised humans cannot respond to moral reasons seems misguided. I suggest the same is true for concerns that the Moral Decision Machine relies on its database.

To summarise, I see no convincing reasons to withhold adequate moral reasons-responsiveness from the Moral Decision Machine. I suggest that the burden of proof lies with arguments that artificial agents cannot be adequately moral reasons-responsive. These arguments should explain why the Moral Decision Machine cannot adequately respond to moral reasons as their arguments must claim. If the Moral Decision Machine is adequately moral reasons-responsive, then either the purportedly necessary metaphysical qualities skeptical arguments emphasise (emotion, evolutionary history, moral judgement, etc.) can be possessed by artificial agents, or they are unnecessary for adequate moral reasons-responsiveness.

Supporters of these arguments may concede this but argue that this is no problem for them because the Moral Decision Machine is impossible. I turn to these kinds of objections now.

Objection 6. Historical situations cannot be properly compared to present (or future) situations. This is demonstrated clearest by historical developments – The Moral Decision Machine has a different historical context to the humans in its snapshots and that is sure to involve different contextual empirical facts.

Different contextual facts could make a serious difference to the Moral Decision Machine's performance. Technological advances may cause old strategies to be outdated. Moral scientific progress may uncover previously unnoticed moral reasons. New institutions or social organisations may lead to unprecedented political climates, legal systems, or social norms. In some cases, this may lead the Moral Decision Machine to fail to respond to certain moral reasons. However, it would still be likely to be adequately moral reasons-responsive unless there were paradigmatic shifts in the moral reasons humans respond to. The Moral Decision Machine, like a human time-traveller, would still be adequately moral reasons responsive even if its moral reasons were outdated at times.

It is unlikely that these contextual differences would cause the Moral Decision Machine to be unable to function at all. First, while some moral reasons may be reframed over time, the moral reasons that the humans in the database responded to would presumably still be

present. Second, some decisions about the *relevant* features of situations would need to be made. The empirical situations need to be identical relevant with respect to the facts that ground moral reasons, but not completely identical. Contextual facts that change over time would be unlikely, I suspect, to change the facts that ground moral reasons in every-day situations. This is especially true if the facts are abstract – a 'driver on their phone' may be relevantly identical to 'a distracted driver' in some cases.

All that said, if the world changed dramatically over time in a way that prevented the database from being constructed in the first place, then the Moral Decision Machine may be unable to be created.

Objection 7. The probability of empirically identical situations occurring is vanishingly small, and therefore the database of moral decisions must be impossibly large for The Moral Decision Machine to work. Perhaps no two empirical situations can ever be identical as a matter of conceptual truth.

This objection is like the previous one. Again, the emphasis here should be on the relevant empirical facts rather than whether the snapshot is completely empirically identical to the Moral Decision Machine's situation. There may be some irrelevant empirical differences that can be overlooked or abstracted away.

In fact, you should not overestimate the degree to which these moral reasons track empirical changes. Humans respond to moral reasons similarly in many different empirical contexts. Most humans, for example, think that killing humans is wrong in almost all circumstances. So, at least for humans there are not all that many relevant empirical facts involved in that kind of moral reason.

The Moral Decision Machine needs a margin of error for empirical identicality. That margin of error will need to be calibrated and refined. It should be high enough that the database is a manageable size, but low enough that most people agree that all the relevant differences in moral facts are accounted for. This would be difficult to achieve but seems to be possible.

Objection 8. It is impossible to create detailed enough snapshots of the empirical situations for the database to function. To trawl the database in a reasonable time frame, even assuming extremely high computational power, will always require compression and resultingly the loss of empirical information, which may distort the supervening moral reasons.

The issue here is twofold. One is that moral reasons track minute changes in empirical situations that would be distorted by compression. To which the reply is that moral reasons do not generally track the smallest empirical facts, and that compressed snapshots are therefore unlikely to distort the moral reasons contained in them. The second is that the database needs to be small enough for the Moral Decision Machine to find the relevantly identical snapshot in a reasonable time. As I suggested, calibrating the margin of error in identicality will help this. But I also want to draw attention to the success of artificial systems in performing these kinds of tasks – image recognition algorithms and large language models are two effective contemporary examples that trawl through large databases to find accurate matches in good time.

The claim that the Moral Decision Machine is practically impossible seems more convincing than the claim that it fails to adequately respond to moral reasons. However, I have offered an initial defence of its possibility. Central to this has been the idea that the snapshots do not need to be completely identical to the Moral Decision Machine's situation, but only relevantly identical. If being 'relevantly identical' does not involve an overwhelming number of empirical facts (and we have reason to believe, based on the relative robustness of human judgements about the moral reasons in situations, that this is so), then the Moral Decision Machine has a good claim to being practically possible.

6.4. The Problem with Moral Deference

I have argued that the Moral Decision Machine is adequately moral reasons-responsive. The rest of the chapter argues that, despite this, the Moral Decision Machine is not a moral agent. I will argue that it is not a moral agent because it is dependent on moral deference and morally deferential agents cannot be responsible. I will then argue that this is not limited to the Moral Decision Machine, and that many artificial systems cannot be moral agents because they are morally deferential.

In moral epistemology, many think that moral deference is problematic. They think moral deference signifies either a lack of moral understanding or unjustified reliance on moral authorities. The argument for this stems from a simple intuition that some cases of moral testimony are undesirable.

Here's Alison Hills' example of where moral testimony feels wrong:

> Eleanor has always enjoyed eating meat but has recently realized that it raises some moral issues. Rather than thinking further about these, however, she talks to a friend, who tells her that eating meat is wrong. Eleanor knows that her friend is normally trustworthy and reliable, so she believes her and accepts that eating meat is wrong.
>
> Many people believe that there are strong reasons not to form moral beliefs on the say-so of others, as Eleanor does. I will call these people "pessimists" about moral testimony. (Hills, 2009)

Pessimism is relatively popular. Andreas Mogensen writes: "[P]essimists have argued convincingly that this is the case. The key issue isn't whether our intuitions accord with pessimism, but why." (2017)

Explaining the attractiveness of pessimism about moral testimony requires some clarifications. First, pessimists do not target *all* cases of testimony. They target *some* cases of *moral*[53] testimony. Testimony about empirical facts, children forming moral beliefs based on testimony, or taking on moral advice (providing you think about it yourself too) all typically satisfy pessimists. The cases they do target, like the Eleanor case, are those in which testimony is taken to be sufficient and definitive evidence for a moral belief or action. This type of reliance on testimony is 'moral deference'.

The reason why moral deference is problematic is debated. Hills (2009) offers the standard explanation: moral deference is problematic because morally deferential agents do not demonstrate the necessary level of moral understanding. "Moral understanding involves a grasp of moral reasons, or more precisely, a grasp of the connections between moral reasons and moral conclusions" (Hills, 2020, p. 408) An agent may be unable to adequately respond to a moral reason they came to believe through deference if they fail to understand that moral reason. For example, an agent may defer in their belief that 'eating meat is wrong' without moral understanding, and their lack of understanding may lead them to fail to believe that 'eating chicken is wrong'.

But moral understanding accounts are not the only game in town. A second explanation for why moral deference is problematic turns on autonomy and character. Robert Howell (2014) suggests that understanding based accounts are flawed, and that the problem with moral

[53] With the possible exception of aesthetics, which, as Mogensen (2017) pointed out, produces some of the same intuitions about testimonial evidence.

deference is that the moral beliefs formed are not consistent with the character and identity of the agent. Mogensen (2017) agrees with Howell's criticisms of understanding based accounts but suggests that the problem is that moral deference undermines the authenticity of the agent. He says, "To be authentic, the beliefs which guide us through life must give expression to the true self. This seems to require that we should decide moral questions on our own terms, so far as we can, so that our own moral sensibility is manifest in the values and ideals by which we live. By contrast, relying on moral testimony puts us in a condition of inauthenticity, since the moral beliefs that guide us fail to give expression to the traits that make us who we are, deep down." (Mogensen, 2017, 277). Mogensen's concept of authenticity is also called 'autonomy' by some of his sources. I use that terminology here, to cohere with the rest of the thesis (especially the next chapter). Furthermore, this concept of autonomy is also well-known as a condition for *responsibility*. In fact, Mogensen's concept of authenticity seems to closely resemble Harry Frankfurt's (1988). The natural upshot, as I will discuss a little more later, is that if reliance on deference does undermine autonomy, then morally deferential agents cannot be (fully) responsible for their deferential beliefs and actions.

Others think that there are *epistemic* problems with deference. That is, there is nothing wrong with deference per se, but that agents are not typically well-informed enough to morally defer properly. Sarah McGrath (2009) offers the explanation that moral deference is problematic because there are formidable epistemic difficulties in identifying a person with superior moral judgement. Paulina Sliwa (2012) similarly suggests that it is practically impossible to identify a moral authority. On these accounts, moral deference is rarely justified, but otherwise not more problematic as a source of belief than other forms of deference.

None of these explanations need, it seems to me, to compete with one another. It may be that moral deference *both* implies a lack of moral understanding *and* undermines autonomy, while normally being epistemically unjustified. However, the important explanation in the following section will be that moral deference is *autonomy undermining.* Because that is the relevant property for the Moral Decision Machine's moral agency. Most moral understanding and epistemic accounts, I will argue, ought to claim that the Moral Decision Machine's deference is benign. However, the autonomy explanation about moral deference leads to the conclusion that the Moral Decision Machine cannot be autonomous and therefore cannot be a moral agent.

6.5. Artificial Moral Deference

The Moral Decision Machine forms moral beliefs and acts upon them based on the say-so of the human in the snapshot it refers to. It does not explicitly defer in the sense that the human in the snapshot utters "I act in this way because acting in this way responds to *this* true moral reason" or "*this* is right" and the Moral Decision Machine takes this to be definitive evidence for the truth of that. But nonetheless its moral actions are performed through the assumption that the action of the human is definitive evidence for belief. This seems an example of the kind of trust distinctive to deference.

> The strongest type of trust, deference, is to believe that p because the speaker has said that p, whatever your other reasons for or against believing that p - and so even if you have a lot of other evidence against it, even if p seems completely crazy to you. You take yourself to have sufficient reason to believe that p, whatever other evidence you have. (Hills, 2020, 402)

The Moral Decision Machine acts based on the human in the snapshot's actions in a deferential way. It acts as if the human's actions are definitive evidence for the rightness of acting in that way, regardless of any other evidence.

We might still wonder whether this is rightly called deference because of two key differences. First, the human does not *intend* to testify; second, the human does not *say* anything. However, on reflection, neither should seem to be problematic. Deference is something *performed by* the deferring agent, not something that requires the agent to consent or intend to be deferred to. Likewise, the fact that deference is often based upon utterances does not entail that it is necessarily based on utterances. It seems quite right that I can defer to a gesture or demonstration. So, despite the Moral Decision Machine deferring to demonstrations that aren't intended to be deferred to, it still performs deference. Since it defers on moral issues specifically, then it performs *moral* deference.

So, given that the Moral Decision Machine morally defers, is this deference problematic? The moral understanding explanation for why moral deference is problematic is that it interferes with moral reasons-responsiveness. Under this explanation, the deferential nature of the

Moral Decision Machine is unproblematic. Despite being morally deferential, it *does* adequately respond to moral reasons, and therefore *does* have moral understanding; so under this account the Moral Decision Machine's moral deference is benign.

How about the epistemic explanation? Is it bad practice, on epistemic grounds, for the Moral Decision Machine to defer? It seems not. The humans being deferred to *are* moral experts compared to the Moral Decision Machine. The Moral Decision Machine is 'morally blind', it has no capacity to respond to moral reasons *other than* moral deference. So, it seems justified on epistemic grounds for the Moral Decision Machine to rely on moral deference, as it is for a blind person to rely on directions.

The other explanation for why moral deference is problematic is that it is autonomy-undermining. In this case, the Moral Decision Machine's moral deference *does* seem problematically autonomy-undermining. The Moral Decision Machine's actions are not based on *its own* capacities or values. I will discuss artificial systems' autonomy and its relation to their moral agency further in the next chapter. For now, there remains a question mark hanging over morally deferring agents' autonomy, and therefore their moral agency. At least, if moral deference *does* undermine autonomy, then it seems right to think that the Moral Decision Machine is not autonomous and therefore not responsible. Thus, while the possibility of the Moral Decision Machine might be good news for a would-be designer of artificial moral agents, it also poses a new challenge: can an artificial agent be adequately responsive to moral reasons *without* deferring? That is, can an artificial agent be adequately responsive to moral reasons while being *autonomous?*

One further point is worth discussing here. Recall that pessimists think there is something particular to *moral* deference that is problematic, rather than all cases of deference or all reliance on testimony. If moral deference is autonomy-undermining, is it *especially* or *uniquely* autonomy-undermining compared to non-moral deference? Mogensen (2017) suggests it may be something to do with the centrality of moral beliefs to an agent's character but leaves it as an undecided issue.[54]

[54] I want to avoid metaethics in general, but here, I think the right answer might demand a metaethical response. Perhaps there is something different about moral reasons compared to other types of reasons that makes moral deference autonomy undermining. For example, if moral facts were mind dependent, it would make more sense that they were autonomy undermining to defer upon, because an agent would sacrifice their ability to *make*

The Moral Decision Machine is dependent on moral deference. But are most other artificial systems? Many, I think, are. Artificial systems that are directly controlled by others, learn under 'supervision', or motivate their actions by analysing human-generated data might all be said to defer. Artificial systems that are directly controlled by others are a clear case and can be held to defer even more straightforwardly than the Moral Decision Machine. Artificial systems that learn through 'supervision', that is, artificial systems that are calibrated and trained by human supervisors actively offering positive and negative feedback, also defer – they respond to moral reasons that are they take to be true purely based on the supervisor's feedback. Finally, artificial systems that act based on human-generated data can also be said to defer. Although this is potentially a borderline case. Artificial systems that work from databases of text (like large language models) or statistical data about human behaviour (which is similar to the Moral Decision Machine) might be said to defer because they respond to reasons purely based on others actions. There are many missing details here, but it is at least plausible that many artificial systems generally defer to others. If so, then this deference would be generally unproblematic, but if the autonomy explanation for moral deference being problematic is right, it entails that they cannot be autonomous, and thus, cannot be responsible.

One final and more fundamental way in which all artificial systems might be said to defer is through being *designed*. By being designed, they take the moral reasons that they were designed to respond to as definitively true *because* they were designed in this way. Design, and its relationship with autonomy is the central point of discussion for the next chapter.

6.6. Conclusion

I considered the arguments that artificial systems cannot have adequately respond to moral reasons because they lack moral judgement (the anti-codifiability argument) or emotions, consciousness, evolutionary history, or some other metaphysical property. I argued against these positions with a counterexample: the Moral Decision Machine. I replied to potential

---

truths in some way. (Though I prefer a 'moral error' approach, in which moral facts *seem* mind dependent because we do not know them yet. This would also seem to support the autonomy-undermining nature of deference because moral facts would be mind-dependent *in practice*.)

concerns about whether the Moral Decision Machine truly adequately responds to moral reasons and found no reason to deny this. Then I considered concerns about whether the Moral Decision Machine is possible, and argued that, given certain plausible assumptions, it does seem possible. The possibility of the Moral Decision Machine is an effective counterexample. Arguments that artificial systems cannot adequately respond to moral reasons because they lack certain metaphysical qualities like emotions or a biological ontology fail because the Moral Decision Machine shows that adequately responding to moral reasons requires no such specific properties. The Moral Decision Machine either has the capacity for moral judgement, in which case the anti-codifiability argument is flawed, or it does not, in which case an agent can be adequately moral reasons-responsive without the capacity for moral judgement.

I then turned to another reason for denying the Moral Decision Machine moral agency: it performed moral deference. I discussed various explanations for why moral deference is problematic, including the claim that moral deference is autonomy-undermining. Then, I argued that the Moral Decision Machine *does* commit moral deference, and, if moral deference is autonomy undermining then the Moral Decision Machine is not autonomous and therefore cannot be a moral agent. I suggested that many artificial systems may be said to defer and may be unable to be responsible. Finally, I suggested that being designed might be a universal feature of artificial systems that leads them to systematically defer. A close relation of this idea, that design is autonomy-undermining, is the topic of the next chapter.

7. Can Moral Agents be Designed?

7.1. Introduction

This chapter introduces and evaluates what I call the 'Design Hypothesis'. According to the Design Hypothesis, design is autonomy-undermining and so designed agents are unable to be responsible[55] (Hakli & Mäkelä, 2016, 2019). I will argue that the Design Hypothesis is false; but that a similar hypothesis is true: designed agents cannot be autonomous unless they are designed *in a particular way*.

In 7.2, I outline the Design Hypothesis. It is argued for with an analogy between manipulation, which undermines autonomy, and design. Autonomy is necessary for responsibility, so, if design is autonomy-undermining then designed agents cannot be responsible.

In 7.3, I outline a definition of design: to design something is to act in a way that confers a function to that thing by modifying or creating its physical material (Borgo et al., 2014). Then, I discuss design's scope of influence – arguing that if design is like manipulation, then it is like *global* manipulation (McKenna, 2004) and undermines a designed agent's ability to be responsible.

Section 7.4 discusses approaches to the conditions of responsibility. The Design Hypothesis assumes a historical account of responsibility (E.g., Fischer & Ravizza, 1998; Haji, 1998; McKenna, 2016; Mele, 1995) whereby an agent needs to meet historical conditions, such as being free from manipulation, to be autonomous and responsible. I outline compatibilist historical and non-historical[56] approaches to responsibility and present the argument that nonhistorical conditions cannot explain manipulation cases.

Section 7.5 presents some design cases that resemble manipulation. I discuss two cases from the responsibility literature (Derk Pereboom's (2001) four case argument and Alfred Mele's (2008) zygote case) and one from artificial systems ethics (artificial servants (Chomanski,

[55] Per the previous chapter, design might be seen to as a special type of deference. But I will not discuss this further.
[56] Supported by the likes of Harry Frankfurt (1988), Gary Watson (1999) and Susan Wolf (1987),

2019; Musiał, 2022; Walker, 2006)). The section argues that, based on the similarities between these design and manipulation cases, design can undermine autonomy.

Section 7.6 considers some challenges for historical accounts' explanations of design and manipulation cases. This is the '*no difference'* reply. It suggests that if a manipulated agent is nonhistorically identical to a responsible agent then the manipulated agent is responsible. A similar reply has been offered for design cases. The 'no difference' reply for design cases is that the designed agents are relevantly historically *and* nonhistorically identical to responsible agents, and so can be responsible. Some think the no difference reply is convincing. To answer it, a historicist needs a well-defined historical condition of autonomy that can identify meaningful differences between responsible agents and designed/manipulated agents.

In 7.7 I narrow in on an historical condition of responsibility that fits the bill. Drawing on recent proposals (Deery & Nahmias, 2017; Waller, 2014), I define the historical condition of autonomy as the following: for an agent to be autonomous over an action, that action's 'causal source' cannot be any other agents' action(s). Where a 'causal source' is a causal variable that "bears the strongest causal invariance relation to [the manipulated agent's action] among all the prior causal variables … that bear such relationships" (Deery & Nahmias, 2017). In coming to this definition, I argue that an agent need not *intend* to undermine autonomy to perform an autonomy-undermining action. Using this definition, a historicist can explain that there *is* a difference between designed/manipulated agents and responsible agents like humans – and therefore reject the no difference reply.

Finally, section 7.8 argues that according to this historical condition the Design Hypothesis is false. But a similar hypothesis is true: design is autonomy-undermining when designer's actions are the causal source of the designed agents' action(s). In most cases, then, design is autonomy undermining. However, it is possible to design agents whose actions' causal sources are natural, physical, or otherwise non-agential forces. Agents designed like this would not have their autonomy undermined, and therefore could be responsible. I consider some options for achieving this, suggesting that while both random design and evolutionary design could be autonomy-conserving, designing agents whose actions' causal sources are evolutionary forces seems more achievable.

7.2. The Design Hypothesis

Raul Hakli & Pekka Mäkelä (2016, 2019) seem to endorse[57] the Design Hypothesis in the following:

> [E]ven though robots could be programmed and engineered to have all the capacities required of moral agents, they would still not be moral agents, because such programming and engineering is closer to […] autonomy-undermining manipulation […] It is precisely the fact that the responsibility-relevant property X was engineered that undermines the responsibility attribution to the agent: Robots cannot be morally responsible because they are designed and programmed by other agents to have the "character" they have. (Hakli & Mäkelä, 2016)

Hakli & Mäkelä borrow the phrase 'autonomy-undermining' from Alfred Mele (1995). The sense of autonomy being used should be distinguished from engineering 'autonomy'. Engineers and computer scientists normally use a technical definition of autonomy where an 'autonomous' agent is an agent that acts independently from direct control by another (See Noorman & Johnson, 2014). This has entered popular parlance; but 'autonomous' drones and vehicles are *only* autonomous in this engineering sense (I call this 'engineering-autonomous'). The autonomy Hakli & Mäkelä and Mele refer to is different. It is 'philosophical' autonomy, where an autonomous agent has 'self-rule' or 'self-determination' over their actions . This is the sense of autonomy I am using.

That autonomy is necessary for responsibility is generally accepted, although there are terminological discrepancies. Martin Fischer and Mark Ravizza (1998) call this condition of responsibility 'guidance control', others refer to it as a 'freedom' condition. There is typically little functional difference between the terms in the literature (Though some argue for a greater difference e.g., Fischer, 2017). I take it that any agent that is sufficiently autonomous to be responsible also satisfies conditions for 'guidance control' and 'freedom'.

Responsibility is standardly said to have 'control' and 'epistemic' conditions. That is, to be responsible for an action an agent needs to be *in control* of the performance of the action and

[57] There are several potential sympathisers with the Design Hypothesis. In Military Robotics, some think that artificial systems cannot be autonomous (Hellström, 2013; Schulzke, 2013). Design seems to play *some* element in this. Similarly, others in machine ethics suggest that artificial agents cannot be autonomous because they are programmed by designers (Bringsjord, 2008; Grodzinsky et al., 2008; D. G. Johnson, 2006; Torrance, 2008). Chomanski (2019) and Musial (2022) claim that designing agents is *manipulative* and will show up later.

needs to be *aware* of the action's consequences. The focus here is on the 'control' side of things. In those terms, an agent with insufficient autonomy fails to be relevantly in control and an agent with sufficient autonomy is relevantly in control. Henceforth, reference to 'autonomy' or meeting the 'conditions for autonomy' refers to the level of autonomy sufficient for responsibility.

Intuitively, engineering-autonomous drones and vehicles are expected to lack this autonomy. This seems likely. Engineering-autonomous vehicles, for example, do not autonomously follow the laws of the road or select their destinations. Nor do they have the general ability to autonomously select their goals or rules. Consequently, engineering-autonomous drones and vehicles cannot be responsible. Which is unsurprising because no-one expects them to be held responsible. This is not evidence for the Design Hypothesis, of course, but merely an example of how some agents do not meet the conditions for autonomy. There is not (yet) any reason to think that *being designed* is what makes engineering-autonomous vehicles fail to meet the conditions for autonomy.

The Design Hypothesis can, then, be stated as the following:

Design Hypothesis: The act of design always causes a designed agent to fail the conditions for autonomy needed to be responsible.

The Design Hypothesis assumes autonomy has a historical condition. That is, it assumes that an agent needs to have the right kind of history to be autonomous (and thus to be responsible). Then, it suggests that design is the wrong kind of history. Later, I will discuss the historical conditions of autonomy and theories of responsibility in more detail. But first, I clarify what is meant by 'design', and the scope of its potentially autonomy-undermining force.

7.3. What is Design?

This section defines design. Clarity in how 'design' is to be understood is essential for evaluating the Design Hypothesis. I then use this definition to make some initial distinctions between designed and undesigned agents, and about the scope of design's influence.

Let us start with Daniel Dennett's (1981) 'design stance'. Adopting the design stance is to explain an objects behaviour by assuming that it has some *function* relative to a designer. As

Vermaas et al. point out (Vermaas et al., 2013), this can mean two things. First, it can mean that an object or agent was *intentionally designed* by an agent. Second, it can mean that an object or agent can be explained in terms of some function – that is, they can be understood to be 'teleologically designed'.

> In the teleological design stance a person *y* predicts the behaviour of an entity *x* by appeal to the assumption that *x* is an entity with a purpose and with parts that have functions that are all assigned by person *y*. In the intentional designer stance a person *y* predicts the behaviour of an entity *x* by appeal to the assumption that *x* is an entity with a purpose and with parts that have functions, and by appeal to the assumption that this purpose of *x* and the functions are assigned by an entity *z* that person *y* describes as a rational agent with certain overarching goals and certain perceptual and behavioural capacities.

Biological agents and natural forces can be explained in terms of teleological design. We might say that the frog's tongue is long because its *function* is to catch flies, despite the frog not being designed. Or we might say that the oceanic tides *function* to clear the beach of trash, despite there being no agent that designed them for this purpose. These examples satisfy teleological, but not intentional, design.

Some might think that teleological design is not design *at all*, but even if so, that should not affect the truth of the Design Hypothesis because the only plausible target of the Design Hypothesis is intentional design. A Design Hypothesis that targets teleological design would clearly be mistaken, because almost all systems can be explained in terms of teleological design and some of those systems are autonomous. Humans can be explained in terms of teleological design, we can ascribe functions to human organs – the heart functions to pump blood, eyes function to enable sight, etc. and subsequently to humans themselves. But humans are not, by that token, less autonomous.

So, the Design Hypothesis must target *intentional design*. Such that the designed object or agent's function is the product of a designing agent's intentional action. Humans, tides, and frogs are not intentionally designed in this sense. But artificial systems are.

Intentional design is a standard approach for explaining the design of artificial systems (which are, in philosophy of technology, sometimes called 'artifacts').

> It is, I think, admitted on all hands that human purposes and intentions have something to do with the functions of artifacts. But a fairly common view is that artifact functions are directly and exhaustively determined by individual and/or collective human intentions. (Preston, 2009, p. 218)

> The function of an artifact is derivative from the purpose of some agent in making or appropriating the object; it is conferred on the object by the desires and beliefs of an agent. No agent, no purpose, no function. (McLaughlin, 2000, p. 60)

Generally, designers' intentions are taken to confer functions to designed objects and to involve the selection of the physical constituent of the designed object. This is nicely captured by Borgo et al.'s 'ontological definition' of design:

> An artifact A is a physical object which an agent (or group of agents) creates by two, possibly concurrent, intentional acts: the selection of a material entity (as the only constituent of A) and the attribution to A of a quality. (Borgo et al., 2014)

In my view, design plausibly includes at least *modification* of the material entity. That is, picking up a stick and using it to scratch your back is not *designing* a backscratcher. But attaching a rough sponge to the end of the stick to the same end is. This is a matter of contention in the literature, some (like Borgo et al.) think that selecting the stick to be a backscratcher is to design it because the stick has a function conferred upon it and is selected because of its ability to perform that function. I will not go into further detail, but I agree with those that think that this is *not* design because design involves modification of the material by the designer.

I adopt the following definition of designing:

The act of design: The act by which an agent (or agents) intentionally modifies or creates a physical material such that the resulting system (the designed system) has a function.

Often, designed systems are used for a different function than intended by the designer. Ruth Millikan (1999) distinguishes between a designed system's 'direct function' (the designers intended function) and its 'derived function' (what it is actually used for). For the purposes of outlining design here[58], it can be accepted that designed systems' derived functions sometimes replace their direct functions[59]. The important thing is that the act of design confers the direct function initially. Talk of designed systems' functions is henceforth talk of their direct function.

[58] While it is sufficient to understand design functions like this here, the connection between changes in derived functions over time and the manipulation of designed agents seems worthy of further exploration. For example, it might be seen as manipulative to assign the derived function of being a servant to a designed agent with the function of being free.

[59] A recent example is Ozempic (semaglutide), which was originally used to treat diabetes, but is now often used to treat obesity.

Three further clarifying points. First, a designed system's function can be broad and abstract. One might design a material to be *strong* or to *conduct heat* without having a more specific function in mind. Second, an act of design can *fail* if the designed system cannot perform its function – such as designing a bridge that cannot support the intended weight. Third, designing *agents* is not a special case. Agents can be designed to perform functions and their materials can be created or modified for that purpose. Designing a software agent is an obvious case of design.[60]

Being designed may, on one interpretation, be what makes a system 'artificial'. But even if you do not take that conceptual route, under this definition every artificial system must be designed. You might wonder whether even our most advanced, independent, and complex machine learning systems are designed. I think that this must be the case – and that, furthermore all possible artificial systems must be designed in the sense just outlined. If in the future we could make artificial systems that were functionally equivalent to humans, they would still be designed. Their function would be, roughly, to be functionally equivalent to humans, and designers would act to cause their material constituent to perform this function. The same kind of reasoning applies to multi-functional systems like advanced robotics, generative AI, and classification algorithms. All are materially modified by designers to perform their function, whatever that should be.

Since agents can be designed, you might wonder whether humans are. You might think that the human species was designed by evolution or God (as suggested in Danaher, 2020). First, evolution (or, more precisely, evolutionary forces) cannot be an intentional designer. If it designs agents at all, it does so in the sense of 'teleological design' but not 'intentional design'. It is a *force* or *law*, not an agent[61] that can perform actions. The same applies to physical forces like gravity or entropy. Things like humans, frogs, and hurricanes are not designed in the sense in which I am using it.

As for design by God, I assume that humans evolved over millennia and that creationism is false. However, you could imagine that some agent designed the universe or natural laws, and

---

[60] Software agents fulfil the material modification requirements for being designed even if they are designed digitally because the software is physically embodied in computer hardware which is physically modified by the design process.

[61] See 3.2. for a definition of agency, which supports this because natural forces are not meaningfully situated within an environment.

thus designed the human species. One possibility is to think of God as more *force-like* than *agent-like*[62], and that, like evolutionary forces, God cannot perform *actions* and therefore cannot be a designer. However, some might think that God must be an agent – I disagree with this. But even if true, this would not be a counterexample to the arguments of the chapter. My account of autonomy in 7.8. suggests that even if God was an agent and conferred a function upon humans through some distal physical modification, this would not be autonomy undermining design.

The Design Hypothesis is that design undermines a designed agent's autonomy. On the understanding of design just outlined, the intuitive force of this is, I hope, clear. A designed agent's autonomy is worth questioning because their function was determined by another agent. That said, I will argue that the Design Hypothesis is false, some designed agents can be autonomous.

Having defined design and discussed some example cases, I move on to the scope of design's influence. The Design Hypothesis claims that designed agents cannot be responsible in general, rather than cannot be responsible for a specific act or disposition. The relevant distinction here is that an agent can fail to be autonomous 'locally', i.e., for a small proportion of actions, in which case they can still be responsible in general. Or they can fail to be autonomous 'globally', i.e., for a large proportion of actions, in which case the agent is not responsible. For the Design Hypothesis to be on the right track, design should be plausibly globally autonomy undermining.

This tracks McKenna's (2004) distinction between global and local manipulation. An action is 'locally' manipulative when it only determines an agent's action in certain circumstances. I am globally manipulated when manipulation affects a large amount of my beliefs or values, such that my entire decision-making process is altered. In global manipulation cases, the manipulated agent is not responsible.

So, if design is autonomy-undermining, is it globally or locally autonomy-undermining? There are some acts that seem to satisfy my definition but only locally undermine autonomy. Consider the following example. I genetically engineer a human zygote by altering a single

[62] While I am entirely uneducated in religious matters, I take it that God is, like natural forces, neither embodied nor physical and thus operates on the wrong sort of metaphysical level to be considered a normal agent with intentions that can perform actions.

gene – I change the gene intending to increase the cancer-immunity of the resulting human. I declare "I have designed this human to be cancer resistant!"[63] This seems like intentional design – it is an intentional physical modification that seems to confer a function. It also seems to be, if anything, only locally autonomy-undermining (and perhaps not autonomy undermining at all). However, it does not, I think, really confer a function. As mentioned, evolutionary forces confer functions through teleological design. Intentionally designing an agent should confer a function that can be used to effectively explain the actions and behaviour of the agent. This is unlikely to be true for the cancer resistant human because the explanatorily effective interpretation of their function is still the evolutionary, teleologically designed function. So, the cancer resistant human is not designed.

Why adopt this restrictive account of conferring functions? My reasoning draws on the distinction between 'derived' and 'proper' functions. A system gains a 'derived' function when it *effectively explains* a reasonable amount of the system's use or performance. Adapting systems for idiosyncratic use does *not* confer a derived function unless that derived function is an effective explanation of much of the system's use or performance. Likewise, modifying zygotes by genetically engineering them does not confer a function unless that function is reasonably explanatorily effective for much of the resulting human's behaviour and functioning. Theories of design that focus on 'use-plans' (Borgo et al., 2014; Vermaas & Houkes, 2006) take a similar position – a designed object must have a 'plan' that explains what it is used for – that is, the function conferred must be sufficiently explanatorily effective.

A human could still be designed, I think, but it would take a more drastic modification than altering a single gene to confer a sufficiently explanatorily effective function. Dogs, it seems, through a long process of genetic engineering via selective breeding, *were* designed. Their design offers a reasonable explanation of most of their behaviour. Though this is perhaps a borderline case, the conditions are, I hope, clear. In contrast, creating artificial systems is unmistakably design because it confers their *only* function.

[63] Declaring that you have conferred a function is not necessary to intentionally design something but is an option that increases dramatic effect.

So, Design must confer the either the *only* function of a designed object, or the designed function should be reasonably explanatorily effective. If so, then design is globally autonomy-undermining if autonomy undermining at all.

Note that if you would like to adopt a broader version of design such that the cancer resistant human *is* designed, then the Design Hypothesis will be immediately false (since the cancer resistant human is responsible). However, my argument can proceed by replacing the Design Hypothesis with this narrower version: 'An act of design that confers a reasonably explanatorily effective function undermines autonomy such that the designed agent is not responsible'.

7.4. Historical and Nonhistorical Accounts of Responsibility

The Design Hypothesis is that design is autonomy undermining and undermines responsibility. For design to be autonomy-undermining, autonomy must be history-sensitive. It should make an autonomy-relevant difference whether among two functionally identical agents one agent is designed and the other is not. This issue, and others like it, are discussed in detail in the literature on responsibility and free will. Exploration of these accounts plays a key role in my arguments about the Design Hypothesis. There are three approaches to setting out the conditions for autonomy: historical compatibilism, nonhistorical compatibilism, and incompatibilism. In this section, I explain the difference between these approaches and present the cases that led some to prefer historical compatibilism.

Incompatibilists believe that you can only be free, autonomous, and responsible if determinism[64] is false. They think that to be responsible an agent must have 'freedom' or 'the ability to do otherwise' in the sense of being able to act indeterministically. I confess I find incompatibilism deeply unintuitive, but, as an olive branch, I suspect incompatibilists can accept many of the conditions for autonomy about to be discussed anyway[65]. In any case, I will focus on compatibilist accounts.

---

[64] The claim that there is only one causal outcome of any physical situation.

[65] Is there an incompatibilist free will condition of moral agency? It's not a popular view, especially in machine ethics, though Deborah Johnson (2006) seems to support it. Partly in the interest of space, I assume that most

Compatibilists believe that even if determinism is true, you can be free, autonomous, and responsible. Semicompatibilists think that if determinism is true, you can be responsible and autonomous but not free. Since the focus is on responsibility and autonomy, when discussing compatibilism, I also target semicompatibilism.

I now summarise the competing historical and nonhistorical compatibilist accounts of autonomy. Supporters of historical accounts (historicists) argue that nonhistorical conditions and historical conditions are both necessary and together jointly sufficient for autonomy. Supporters of nonhistorical accounts (nonhistoricists) argue that nonhistorical conditions are necessary *and* sufficient for the autonomy. So, both accounts agree that autonomy contains nonhistorical conditions, and both can share definitions of these conditions. So, I first outline the nonhistorical conditions, before moving on to historical conditions.

There are various accounts of autonomy's nonhistorical conditions. Two prominent ways of defining them are in terms of whether actions fit one's 'real self' (Wolf, 1987) or 'deep self'; and in terms of whether an agent's mechanism for producing action is reasons-responsive (Fischer & Ravizza, 1998). Both approaches are compatibilist: they assert that an agent can be autonomous even if determinism is true.

Real-self accounts hold that to be autonomous, an agent's disposition to act needs to 'fit' or 'mesh' with their higher-order desires or considered values. Traditionally these are represented by Harry Frankfurt's claim that to be an agent needs a 'fitting' relationship between their higher and lower order desires. Gunnar Bjornsson (2016) says that these accounts "standardly understand the deep self as some privileged internal aspect of the agent's psychology, such as higher order attitudes (Frankfurt, 1971), value judgments, plans (Bratman, 1997), or "cares" (Shoemaker, 2003)". The conditions for autonomy in these sorts of accounts are nonhistorical – an agent is autonomous if they have the right relation at a given time between their 'real' or 'deep' self and their actions, regardless of *how* they came to possess their 'real' or 'deep' self.

Reasons-responsive accounts hold that one needs to be responsive to reasons to be autonomous. The most well-known reasons-responsiveness account is Martin Fischer and

---

incompatibilist conditions for moral agency have a compatibilist near-equivalent (See King, 2013 for some arguments along these lines), and that those which do not are marginal views.

Mark Ravizza's (1998) which *does* include a historical condition for autonomy (which, as mentioned, they call 'guidance control'). But other variations may not include a historical condition, such that an agent is autonomous when they respond to the reasons in the appropriate way. R. Jay Wallace (1998) and perhaps Susan Wolf (1993) focus on the ability to respond to *moral* reasons as a condition for responsibility (though this thesis understands that capacity to be distinct from responsibility). For the purposes of this chapter, they reasons-responsiveness accounts of the nonhistorical conditions of autonomy can be taken as a group. They are an alternative to 'deep/real self' accounts but need not include historical conditions.

Whatever the differences between nonhistorical accounts, they are united in claiming that the autonomy only has nonhistorical conditions. Historical accounts disagree. They think that to be autonomous an agent must have (or lack) a certain *history*. Their primary evidence for this is 'manipulation cases', cases in which agents seem to satisfy all the nonhistorical conditions of autonomy while failing to be autonomous.

Manipulation cases were first used as part of the 'manipulation argument' for incompatibilism. Roughly, the argument is that if manipulated agents fail to be autonomous because they were determined to act, then determinism entails that no agent can be autonomous. Some compatibilists think that the tables can be turned: manipulation cases can challenge incompatibilism too. But I do not focus on the success of the manipulation argument here.

Instead, I focus on which compatibilist accounts can better reply to the manipulation argument. Nonhistorical and historical compatibilist accounts compete over who can better explain manipulation cases. In this, historical accounts appear to have a decisive advantage: they can claim that manipulated agents are not autonomous because manipulation violates the historical condition of autonomy, not because being determined violates autonomy. I will now explain this further.

Mele's 'Ann/Beth' case is probably the most well-known manipulation case. My summary of the Ann/Beth case follows.

Ann/Beth: Ann is an ordinary, hard-working philosopher who enjoys working hard and, with the approval of the dean, continues in that manner. Beth is not hard-working, instead enjoys many aspects of her diverse and fulfilling life. The dean is unimpressed with Beth, and, having failed to convince her to change her ways, turns to a team of rogue scientists. The

scientists, through sophisticated new technology, re-jig Beth's psychology (i.e., her values and tendency to endorse values) to be the same as Ann's. The result of this is that Beth becomes a 'psychological twin' of Ann, she not only works hard but enjoys doing so too, and despite being surprised by her new values and dispositions, fully endorses her new lifestyle. (Mele, 1995, 145)

Note that the Ann/Beth case is a global manipulation case, so Beth's *general* responsibility is under question, not her responsibility for any particular action.

Let us suppose (Contra Mele's intentions but corresponding with McKenna's (2012) interpretation) that Beth is sufficiently like Ann that they equally meet the nonhistorical conditions of autonomy. Take it that Ann and Beth are equally reasons-responsive, equally appropriately endorse their values, etc. Ann, as an ordinary human, can be assumed to be autonomous, so is Beth, who is so similar to Ann, also autonomous? Many, like Mele, feel that Beth is *not* autonomous and thus cannot be responsible.

Compatibilist accounts offer different explanations for manipulation cases. Mckenna outlines two types: 'soft-line' and 'hard-line' replies. Soft-line replies claim that manipulated agents are not autonomous. Hard-line replies claim that manipulated agents are autonomous. Historicists take a soft-line reply. They suggest that manipulated agents are not autonomous because they have the wrong sort of history. As Fischer puts it:

> The intuition is simple. The mechanism that issues in behavior (or, more broadly, the way the behavior is produced) can be reasons-responsive, but this sensitivity, or significant features of it, could have been induced externally (by clandestine manipulation, hypnosis, subliminal advertising, brainwashing, and so forth). So reasons-sensitivity is not enough for moral responsibility. *The reasons-responsiveness itself cannot have been put in place in ways that bypass or supercede the agent - the mechanisms that issue in one's behavior must be one's own.* (Fischer, 2004, p. 147, my emphasis)

Mele concludes the Ann/Beth case by suggesting that an historical condition for autonomy is needed.

> Beth's autonomy was violated, we naturally say. And it is difficult not to see her now, in light of all this, as heteronomous to a significant extent. If that perception is correct, then given the psychological similarities between the two agents, the difference in their current status regarding autonomy would seem to lie in how they came to have certain psychological features that they have, hence in something external to their

> here-and-now psychological constitutions. That is, the crucial difference is historical; autonomy is in some way history-bound. (Mele, 1995, 145-146)

Historicists' can thus answer the manipulation argument. They can argue that manipulation argument fails because autonomy has historical conditions and Beth has a different history to Ann. So historical compatibilist accounts can explain why Ann is responsible while Beth is not. As mentioned, this is a soft-line reply: it accepts that Beth is not responsible and argues that there is a relevant, historical, difference between her and Ann.

Nonhistorical accounts cannot adopt this reply because they do not accept that autonomy has historical conditions. Instead, they tend to take the hard-line reply and hold that Beth *is* responsible. Here's Frankfurt and Watson hard-line replies:

> To the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused are irrelevant to the questions of whether he performs his actions freely and is morally responsible for performing them. (Frankfurt 1988: 54)

> For it is hard to see what differences there could be between the natural and purposeful forms of determination that would be relevant to freedom and control. […] If purposeful determination of one's actions by another agent undermines freedom, so does determination by the natural world. (Watson, 1999, p. 361)

> The point here, to which all compatibilists must assent, is that the responsibility-conferring features of the actual processes are identifiable without any further reference to the history (the causal story) regarding the features so identified (Watson, 1999, 364)

Many find hard-line replies of this sort unappealing. "[N]onhistorical theorists are saddled with taking on the jarring claim that agents like Beth act freely and can be morally responsible for what they do." (McKenna, 2016, p. 85). Assuming we do not want to take on this claim, only historical accounts can offer a convincing reply to the manipulation argument. Many compatibilists take this to be a persuasive reason for adopting an historical account of autonomy.

Of course, nonhistorical accounts may have some resources to claim that that the hard-line response is the best (or only viable) response available. This is the kind of argument I discuss in 7.6. But before that, I return to the Design Hypothesis, while assuming that the historical account of autonomy is at least highly plausible.

7.5. Design Cases

Defenders of the Design Hypothesis argue that design is autonomy undermining because it is analogous to manipulation. Some prominent cases in the responsibility literature resemble design much more than the Ann/Beth case, and so can help better flesh out this analogy. In the zygote case, again from Alfred Mele, and one part of Derk Pereboom's 'four case argument', the autonomy-undermining action occurs before the affected agent was ever responsible or autonomous. Acts of design share that timing. I now discuss these two cases with the aim of drawing an analogy between manipulation cases and design cases.

Here is Mele's zygote case.

> Diana creates a zygote Z in Mary. She combines Z' s atoms as she does because she wants a certain event E to occur thirty years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z' s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about E. (Mele 2008: 188)

Zygote Z becomes a person named Ernie who brings about event E. Mele claims that Ernie is 'neither free nor responsible' in general. But Patrick Todd (2013), in a move endorsed by others (Fischer, 2017; Pereboom, 2014) suggests that it is more accurate to say that Ernie is not responsible for valuing A or bringing about E. This difference tracks whether the zygote case is locally or globally autonomy-undermining. Perhaps anticipating this, Mele suggested modifying the story: "As I observed, Diana's assembling Z as she does in Mary—her means of achieving her aim—is a cause of all of Ernie's actions and not merely of his A-ing (by which he brings about E). In a modified version of the story, Diana has a much more extensive aim—to create an agent who performs all of those actions." (Mele, 2008, 190). Later, I will argue that manipulators intentions are not important, however, the difference between the modified and original zygote case are. So, I proceed with this modified version, where A now represents the full set of Ernie's actions and E the consequences of those actions.

Diana's action seems to satisfy the definition of an act of design[66]. She creates the physical material of Ernie with the intention of him bringing about E. Therefore, Ernie is designed, and Ernie's function is to bring about E. Intuitively, Diana's act of design is alike to manipulation in that it causes Ernie to fail autonomy. One conclusion to draw is that if Ernie's autonomy was undermined by Diana's act of design, then acts of design *can* undermine autonomy. This would be an important step for defending the Design Hypothesis, but it is not yet evidence that design is generally analogous to manipulation.

The second case of manipulation-like design comes from the second case of Derk Pereboom's four case argument. It is as follows.

> Case 2: Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic […] and at times strongly so, with the intended consequence that in his current circumstances he is causally determined to engage in the egoistic reasons responsive process of deliberation and to have the set of first and second-order desires that result in his decision to kill White. Plum has the general ability to regulate his actions by moral reasons, but in his circumstances, due to the strongly egoistic nature of his deliberative reasoning, he is causally determined to make his decision to kill. Yet he does not decide as he does because of an irresistible desire. (Pereboom, 2001, pp. 113–114)

The first case of the four-case argument involves manipulation that occurs at the time when Plum (call him Plum1) kills white, so, it is not a design case. But in this second case the scientists intervene before Plum (call him Plum2) is born, so it is closer to design. Like Diana's act in the zygote case, the neuroscientists actions in this case are plausibly acts of design. Plum2's function is to kill White, and he was modified to perform this function. It is plausible furthermore, that this function effectively explains much of his actions because his reasoning is *often* egoistic.[67] This act of design, at least according to Pereboom, causes Plum2 to lack autonomy as Plum2 "is not morally responsible because his action is determined by the neuroscientists' programming, which is beyond his control." (Pereboom, 2001, 114)

---

[66] It is strange, and to my mind false, to consider her an agent (as with God); but I concede her agency by weight of stipulation.

[67] Though without modifying the case further (like the modification Mele suggested for the Zygote case) it might still be an open question whether this egoistic disposition sufficiently explains Plum's actions such that he is *designed*. In the case that he is not designed, then the egoistic disposition must have limited affect, and therefore it seems intuitive that he is autonomous, although he would lack autonomy over and responsibility for killing White.

Pereboom thinks that Plum fails to be autonomous despite the unusual timing of the scientists' intervention. He writes: "Furthermore, it would seem unprincipled to claim that […] Plum is morally responsible because the length of time between the programming and his decision is now great enough. Whether the programming occurs a few seconds before or forty years prior to the action seems irrelevant to the question of his moral responsibility." (Pereboom, 2001, 114)

The differences between the interventions in the zygote case and Pereboom's case 2 are significant. Diana is effectively omniscient, and can entirely predict Ernie's actions, while the neuroscientists are fallible, albeit with some specialised knowledge about Plum's future dispositions. Pereboom's neuroscientists are closer to human designers in this regard. If, as many find intuitive, Plum's autonomy is undermined by the neuroscientists design, that is more evidence for design being autonomy-undermining.

The final case comes from literature about whether we ought to design 'artificial persons'. There is some discussion here about whether designed agents can be autonomous. Walker (2006) argues that designing artificial persons to serve humans would undermine their autonomy. However, Walker seems to think that artificial servants lack autonomy because they are designed *to be servants*, not because they are *designed*.

In addition, Peterson (2011), Chomanski (2019) and Musiał (2022) think that designing artificial servants need not be autonomy undermining. However, they assume a nonhistorical account of autonomy. Of course, if historical conditions for autonomy are denied off the bat, then artificial persons, so long as they can meet the necessary nonhistorical conditions (which they do by stipulation), may be autonomous. But if we reconceive of artificial servants as a potential design case to examine our intuitions about whether design undermines autonomy, the result may be different. As Chomanski notes, artificial servants may not be autonomous under a historical account of autonomy, since "the mental states with which their motivations cohere, or their motivations themselves, appear to come from an external source (from the programmers) rather than from the AIs themselves".

So artificial servants do seem to resemble design cases after all. Furthermore, Chomanski and Musiał argue that design is manipulative by drawing on an Aristotelian conception of manipulation wherein being 'manipulative' is limiting another's choices without due regard for the agent's rational capacities. But the explanation of manipulativeness closely resembles putatively autonomy-undermining design cases.

> To see why programming AIs to be servants is manipulative, consider the following points. AI servants are artificial people programmed to find the deepest fulfillment in attending to human needs. By programming AIs with overwhelming desires to be servants, their programmers are set on making it close to impossibly psychologically difficult for the AIs to act in any way other than to pursue the life-plan that they are given: even a determined AI servant would likely not pick a different career path. This might be so even if they are programmed with the capacity to reflect on their own desires. Upon being told to build an AI servant, the programmers cannot help but suppose that the servant's decision about what life plane to pursue are for the programmers to make.
>
> The programmers position themselves so as to be able to steer the choices of AI servants with a high degree of efficacy, by constructing the AIs' psychology in the way just specified. They orchestrate the AIs' choice space in such a way that only a very narrow range of options is realistically available to them. The AIs' number of options, when it comes to career choice, is reduced to basically one. The AIs' own capacity to respond to reasons, or to engage in reasoning, is ignored. The act of thus limiting another's options and disregarding their rational capacities is expressive of manipulativeness on the programmers' part—that is, of being willing to orchestrate another's choices in an excessive manner. (Chomanski, 2019)

It seems to me that this is just another design case. Under a historical account of autonomy, the process of designing artificial servants seems to undermine their autonomy. At least, if Ernie and Plum2 are not autonomous or responsible, then there seem no relevant differences in the case of artificial servants. If so, then artificial servants are another case of design undermining autonomy. But it does not, of course, show that design *must* undermine autonomy.

However, Chomanski and Musiał both suggest that intentional design is *always* manipulative.

> I believe that designers who choose to intentionally design APs [Artificial Persons] do not— and actually cannot—respect the rational capacities of the APs they design. […] intentionally designing APs is unavoidably manipulative. (Musiał, 2022).

If design is always manipulative, then we have reason to think that design is analogous to manipulation and always undermines autonomy.

The three cases of Ernie, Plum2 and artificial servants offer evidence (but not proof) for the following conditional. If manipulation undermines autonomy by violating the historical condition of autonomy, then design cases may also violate that historical condition because they are relevantly similar.

However, there are two outstanding issues. First, the nonhistorical accounts of autonomy have a response to manipulation and design cases that ought to be addressed. Second, it remains to be seen whether *all* cases of design *must* be autonomy-undermining. The outcomes depend on the final piece of the puzzle: a full explanation of the proposed historical condition for autonomy. If a historical account of autonomy can answer the challenge from nonhistoricists, it will be highly appealing, because it can better explain manipulation cases. If design violates the historical condition of autonomy in that account, then the Design Hypothesis will in a strong position too.

7.6. Qualms about Design Cases

In this section I discuss a central challenge for historical explanations for design and manipulation cases. In the following section I develop an account of the historical condition of autonomy that can rise to it.

Recall that compatibilists can take either a hard-line or a soft-line reply to manipulation cases. Nonhistoricists like Frankfurt and Watson take the hard-line: Ann and Beth are both autonomous because they are relevantly non-historically identical. Historicists like Mele and Fischer take the soft-line: Ann is responsible while Beth is not because they have relevantly different histories.

Equivalent hard-line and soft-line replies are available for design cases. You can take the hard-line and claim that designed agents *are* autonomous because they're relevantly identical to some autonomous agent; or you can take the soft-line and claim that designed agents are not autonomous because they have relevantly different histories to autonomous agents.

To be clear, I support a soft-line reply to design cases. It seems to me that Ernie, Plum2, and artificial servants are *not* responsible or autonomous *because* they were designed. But a significant number of philosophers of responsibility either do not have this intuition or overrule it.

Unsurprisingly, nonhistoricists offer a hard-line reply to design cases too. Harry Frankfurt again:

> We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. … We

> are the sorts of persons we are; and it is what we are, rather than the history of our development, that counts. (Frankfurt, 2002, p. 28)
>
> There is no paradox in the supposition that a [manipulator] might create a morally free agent. (Frankfurt, 1988, p. 54)

Gary Watson thinks that compatibilists must insist "that free agents might indeed be the products of manipulation by designers". He defends a hard-line response to both design and manipulation.

> What seems in the abstract like a responsibility-undermining history might seem so only because it abstracts from the constitutive proper ties of what is supposed to emerge (by design or not) from that history. If we fill out these histories, according to compatibilism, some will be responsibility-undermining, some not. If not, that will be because the design fails to realize some of the [nonhistorical conditions], not because it is deterministic. (Watson, 1999, 365).

But some treat design and manipulation cases differently to each other. Fischer, for example, takes an historicist soft-line reply to manipulation cases but an historicist hard-line reply to design cases.

> Elsewhere, I have argued that the intuition that [Plum] is not in fact morally responsible in Case 1 [in which Plum is manipulated by neuroscientists as an adult] issues from a view that the mechanism that leads to his behavior in that case is not his own (Fischer, 2014). But this view cannot be applied straightforwardly to Case 2; although the neuroscientists preprogram Plum to have the mechanisms of practical reasoning that he has, and to act on them in certain specific circumstances, this does not imply that they are not *his* mechanisms. (Fischer, 2017)

Fischer makes his position clear: he takes a soft-line response to Plum1 – Plum1 is not responsible because he violates a historical condition; and a hard-line response to Plum2 – Plum2's mechanisms *are* his own, so he meets the historical condition and is responsible (and this is the same in the zygote case; Fischer, 2011, 2016). In taking a hard-line response to design cases, Fischer agrees with Frankfurt and Watson about design. They all think that Ernie and Plum2 *are* responsible. Mele expects all compatibilists to adopt this strategy for design cases. He suggests that compatibilists should accept that "when it comes to moral responsibility, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe." (Mele, 2013, p. 180) and that "Ernie should strike them as free and morally responsible (in light of his properties as an agent)." (Mele, 2008, 193).

The hard-line reply to design cases depends on the claim that there is *no relevant difference* between designed agents like Ernie and responsible agents. That is, there is no historical condition for autonomy that can distinguish between Ernie or Plum2 and responsible humans.

Note that Mele and Fischer have a particular reason for adopting a hard-line reply to design cases and a soft-line reply to manipulation cases. In manipulation cases, manipulation causes a relevant historical difference because an agent's beliefs and values are *overridden*. In design cases, there is no overriding and thus, they think, no relevant difference between designed agents and agents 'designed' by mother nature or 'blind forces'. Both result in autonomous mechanisms that are 'one's own'.

Moreover, this is exactly the line that many take against the Design Hypothesis in the literature on artificial moral agency. The hard-line reply to design cases is generally considered to be a standard and effective reply to worries about whether designed artificial systems will fail to be responsible[68]. Behdadi and Munthe summarise the claim as the following.

> [A] condition that excludes machines based on being designed will fail to distinguish between human and artificial entities with regard to moral agency, thus undermining the notion of human beings as paradigmatic moral agents. (Behdadi and Munthe, 2020)

Armed with an understanding of the literature on responsibilities' exploration of the same issue, it should be apparent that this position takes a hard-line reply and is dependent on the truth of the no relevant difference claim.

Hakli and Mäkelä defend the Design Hypothesis with a reply to this, but it seems to be mistargeted. The hard-line reply concerns whether there is a relevant difference between design cases and regular humans. But Hakli and Mäkelä attempt to settle things by appealing to compatibilism.

> Our claim is not that being externally determined undermines autonomy, and mutatis mutandis, moral agency. We are not just giving a rerun to the reasoning that robots are programmed, hence not free, hence not autonomous which has already been discussed in the literature (see, e.g., Bringsjord [2008]) and to which there is a

[68] See (Himma, 2008; 23. Powers, 2013, 235; Peterson, 2011; Musial, 2018; Danaher, 2020; Veliz, 2020;

> standard objection too [sic] (see, e.g., Sullins [2006]): the "paradigmatic moral agents," that is, adult human beings are bound to face the same problem as robots. This is because human beings are what they are due to their genetic programming, upbringing, and contingent influences from their environment. [This 'standard objection' is the hard-line reply]
>
> However, in our view, the possibility that human beings' characters are "determined" by external factors and influences is not the decisive factor in deciding whether an agent is autonomous and responsible. At the background of history-sensitive externalism there is a compatibilist idea according to which autonomy is not necessarily ruled out by determination. Indeed the main point is not about determination per se, rather the relevant issue is the kind of determination: An agent can be autonomous even in a deterministic universe as long as their choices result from their own authentic goals and values. (Hakli & Mäkelä, 2019, 269)

Hakli and Mäkelä here seem to respond to the manipulation argument against compatibilism. They claim that, since compatibilism is true, they may adopt the soft-line reply to design cases. But the challenge, properly understood, need not assume incompatibilism. Compatibilists like Frankfurt think that there are *no* historical conditions of the Autonomy Condition for Responsibility. Compatibilists like Fischer think that there *are* historical conditions of autonomy, but that they are adequately met by design cases. Frankfurt, Fischer and incompatibilists leverage the claim about there being no difference between responsible humans and designed agents, but they cannot be replied to in the same way. Fischer and Frankfurt adopt a hard-line reply while accepting that 'an agent can be autonomous even in a deterministic universe'.

The relevant issue is whether a designed agent can make choices that 'result from their own authentic goals and values'. Fischer and Frankfurt think that they can, while Hakli & Mäkelä think that they cannot. The position in machine ethics summarised by Behdadi and Munthe is best construed as a compatibilist position. Machine ethicists are best interpreted as objecting that any historical condition of autonomy that excludes designed agents will also exclude responsible humans, not that designed agents lack the ability to perform indeterminate action (though Deborah Johnson (2006) is an exception). Hakli and Mäkelä do not speak to this issue.

Hakli & Mäkelä, and any other supporters of the Design Hypothesis, find themselves in need of a soft-line reply to design cases. For that soft-line reply, they must deny the 'no relevant difference' claim. That is, they must offer a historical condition for autonomy that distinguishes between autonomy-undermining design and manipulation histories, on one

hand, and autonomy-promoting regular human histories, on the other. This historical condition, evidently, will not be forthcoming from the usual historical compatibilist suspects like Fischer, Mele, and or even Mckenna (2016) because they support a hard-line reply to design cases. But there are some recent proposals that may yet light the way ahead. In the following section I offer a historical condition that can rise to the task: an agent satisfies autonomy's historical condition for an action if that action's causal source is not another agent's action.

7.7. Historical Conditions for Autonomy and Responsibility

The way forward for the Design Hypothesis is to find a historical condition of autonomy that might facilitate a soft-line response to design cases. The two outstanding issues are whether this historical condition can satisfactorily deny the 'no relevant difference' claim; and what this historical condition entails for the truth of the Design Hypothesis. In this section I outline a historical condition that I think effective, and consider its consequences for designed, manipulated, lucky and evolutionary agents.

Oisín Deery and Eddy Nahmias (2017) offer an 'interventionist' account of causal sources that is intended to facilitate a soft-line reply to design cases. On this account, "[A]n agent can be the causal source of her actions, since often no variable beyond the agent's control will have a stronger causal–explanatory relationship with her actions than relevant variables within her control. On the other hand, the causal source of a manipulated agent's actions lies beyond the agent's control in the intentions of the manipulator." (Deery & Nahmias, 2017)

Using their interventionist model of causal sourcehood[69], Deery & Nahmias claim that actions or events that cause an agent to A with *causal invariance* are autonomy-undermining. To be autonomous, the causally invariant source of an agent's actions must arise from the agent. Being a causal source involves two strong relations: first, the presence and absence of the cause strongly correlates with the presence and absence of the effect; second, there is no alternative causal source that has a stronger relation.

[69] Drawn from James Woodward (2005, 2015)

Actions may have more than one causal source if there are two causes that equally correlate with the effect; in such cases as this, the affected agent may, Deery and Nahmias say, 'share' the responsibility. But these cases are rare, according to Deery and Nahmias' model, most actions only have a single causal source.

Under this model, an agent is the causal source of their own action if their internal mechanisms have the strongest causal relation with their action. In regular cases this is true, you are the causal source of your travelling home if your decision-making process has the strongest causal relationship with that outcome. That is, your deciding *not* to travel home would reliably lead you, in many circumstances, to not travel home, and your deciding *to* travel home reliably leads you to travel home in all sorts of circumstances; furthermore, there is no more reliable cause of your travelling home – there was no reliable kidnapper determined to take you home regardless of your decision, and no antecedent causally powerful manipulation that gave you the disposition to return home. There may be other contributing causal variables, you may decide to return home and then not return home because you get a flat tire, or you may return home because someone asked you to, but these causes have a weaker relationship with the outcome than your decision about returning – they are causes, but not the *causal source*.

Deery and Nahmias (2017) conclude that "[i]t is relevant to an agent's free will and moral responsibility for performing an action, A, whether A has its causal source in the agent (specifically, in a variable representing the output of activity in the agent's [reasoning process that satisfies the nonhistorical condition of autonomy] prior to the agent's A-ing)".

Deery & Nahmias' account convincingly denies the no relevant difference claim for both design and manipulation cases. On their account, designers are the causal source of designed agents' actions and manipulators (when successful) are the causal source of manipulated agents' actions, but autonomous agents are the causal source of their own actions. So, Deery & Nahmias' causal sourcehood historical condition of autonomy results in the claim that designed agents, so long as the designers are the causal source of the designed agents' actions, are neither autonomous nor responsible; while humans, so long as they are not manipulated, indoctrinated, or otherwise act via external causal sources, satisfy the historical condition of autonomy. However, a causal sourcehood condition alone seems to me unsatisfactory because some external causal sources do *not* seem autonomy-undermining.

Hannah Tierney & David Glick (2020) make this complaint against Deery & Nahmias' causal sourcehood condition. In their paper, Tierney & Glick offer several technical challenges about balancing different aspects of causal sourcehood, but, while I do not attempt to do so, those seem resolvable in principle for an historical condition of autonomy. The challenge I focus on is that some causal sources do not undermine autonomy.

> For instance, suppose there is a car accident. There are two causal variables: the presence of a pothole and a driver who is driving in a culpably distracted manner. Both the pothole and the driver driving distractedly are actual causes of the crash—if the driver hadn't been distracted, the crash would not have occurred … and likewise if the pothole wasn't there[.] … It's possible that … [the pothole] would be the causal source of the crash and not the driver, rendering the driver an inappropriate target for responsibility. But surely the driver is morally responsible for the crash. His driving distractedly was an actual cause of the crash" (Tierney & Glick, 2020, edited to remove variables)

In this case Tierney & Glick have the intuition that the pothole does not undermine responsibility or autonomy for the driver, despite the pothole being the causal source of the crash event. I share this intuition.

One explanation for it is that *agents* are different, perhaps more responsibility-attracting, types of causal sources to non-agents. Causal sources that arise from agents perhaps have some kind of multiplier such that even when, strictly speaking, the pothole was the causal source of the accident, the driver remains responsible. Deery & Nahmias (2017) and Marius Usher (2018) think that the right model of causation will have it that agents are more frequently causal sources compared to non-agents. This is because agents' ability to plan and adapt to circumstances means that they cut a straighter line toward particular outcomes than non-agents -- their plans make them stronger and more reliable causal forces. If so, then it seems that Tierney and Glick are pointing to an issue with Deery and Nahmias' model – the *right* model *ought* to have it that the driver *is* the causal source of the crash (that is, if he is indeed responsible for it).

Even if a focus on agents as autonomy-undermining causal sources is not entailed by the model of causation itself, other soft-line replies focus on agents alone as autonomy-undermining influences. Xiaofei Liu (2022) calls these 'another-agent' views; Gabriel De Marco (2023) calls them 'manipulator' views. They argue that the influence of other *agents* is key to the autonomy-undermining force of manipulation and design cases.

This is not a new idea; it first appears in A. J. Ayer's (1963) defence of compatibilism; and reoccurs in Dennett (1984). But I here focus on contemporary examples. Robyn Repko Waller (2014) argues that Ernie fails autonomy's historical condition because Ernie's performance of A is caused by Diana's 'effective intention'. This prioritises agents as autonomy-undermining causal sources, as only agents have intentions.

Waller formulates her general historical condition like so:

> "S is less deserving of blame or praise for A-ing than she would be otherwise if (1) another agent G effectively intends that S A-s, (2) G brings it about that S A-s via intervention, and (3) S did not intentionally bring it about that G intends that S A-s." (Waller, 2014)

On Deery & Nahmias' model of causal invariance, the relevant sense in which G brings it about that S A-s via intervention is when G's action is the causal source of S A-ing. But the salient difference here is that Waller focuses on *agents* specifically – G must also 'effectively intend' that S A-s.

I agree with Liu (2022), Usher (2020), and other supporters of 'another-agent' views (e.g., Herdova, 2021) that focusing on other agents in historical conditions for autonomy can be valuable. However, I want to challenge Waller's focus on the *intention* of other agents. In discussions of the principle of double effect, there are arguments about whether an agent needs to *intend* to cause good consequences to be permitted to incur unintentional but predictable negative side-effects. Suppose that this is true. Suppose that one can *intend* to produce a positive consequence without *intending* to produce negative side effects, even if they are aware of those side effects. Suppose, for example, that Diana intends to manipulate Ernie to perform E, but to do so, she must cause the zygote to be implanted in Mary, Ernie's mother. To be precise, let us stipulate that Diana foresees that Mary will be caused by the zygote implantation to value F and perform B twenty years later. This is not, however, Diana's *intention*. Diana's *intention* is to have Ernie perform A, she does not intend for Mary to perform B, although she knowingly causes it. It seems to me that, despite Diana's lack of intention, she *does* manipulate Mary. Correspondingly, it seems to me that autonomy-undermining interventions in general are independent of designers or manipulators intentions.

Fischer makes the same point: why should *intentions* matter? He argues through a series of cases that the absence or presence of designer's intentions are irrelevant to autonomy and responsibility (Fischer, 2011, 2016). Concluding that "the distal intentions of creators are irrelevant to the subsequent responsibility status of the created individuals, even in a casually

deterministic world." (Fischer, 2016, p. 50). I find his arguments convincing, reference to intentions seems to overcommit. Whether an action is accidental or intentional does not seem to change whether it undermines or conserves autonomy.

However, that does not mean that all agent-focused accounts will fail. One modification to Waller's condition that might do the job to drop the effective intention criterion but keep the agent-focused approach. This is the historical condition of autonomy I propose:

Agentless Sources: S is not autonomous and cannot be responsible for A-ing if another agent G's action is the causal source of S A-ing.

Essentially, I suggest that we keep the Deery & Nahmias' causal interventionist model but insist that the only agents can be autonomy-undermining causal sources. This corresponds with the intuitive results that non-agents like potholes are not autonomy-undermining while agents that unintentionally manipulate are. I think that Agentless Sources can satisfactorily deny the no difference claim, and therefore support a soft-line reply to design and manipulation cases. It also, I will argue, entails that acts of design are autonomy-undermining unless the act of design is specifically performed such that it is not the causal source of designed agents' actions.

The no difference claim is that there is no relevant difference between manipulated/designed agents and autonomous humans. If Agentless Sources is true, then there *is* a difference between design/manipulation and autonomy-conserving histories and the no difference claim is false. But the following objections may arise.

One obvious thought is that if Agentless Sources is true, then humans may not be autonomous. Perhaps many of the causal sources of human actions are other agents. After all, we are social animals from birth, with our upbringing playing a causal role in our future actions; we inhabit a social world as adults, with significant incentives and punishments for conforming or diverging from social norms; finally, we actively interact with and, thus, causally affect one another. All these involve actions by other agents, and any of those might be a causal source of our actions.

There are two reasons why, assuming Agentless Sources is true, ordinary humans are autonomous. First, return to the distinction between global and local autonomy undermining. Ordinary human development plausibly only contains (perhaps several individual instances of) their autonomy being undermined locally. If their autonomy were globally undermined

then it would, like manipulation and design cases, lead humans to not be autonomous or responsible. For example, if a human is indoctrinated from birth by their parents or some nefarious social organisation like a cult or nationalist youth program, then they will probably fail to be autonomous for many actions, and thus fail to be responsible. But an average human, while they might not be autonomous for, say, certain habits or tastes, is autonomous over *enough* of their actions to be generally responsible.

Earlier, I used the modified version of Mele's zygote case just because Diana's intervention in the original case – designing Ernie solely with respect to his performing a single action, was perhaps only locally autonomy undermining. There are many potential variations of Ernie's personality, values, and overall dispositions consistent with his performing a single action, and, assuming that Diana has no interest in those features of him, let us suppose that Diana designs Ernie such that he has the features that he would have had if he had been naturally conceived, and therefore that most of Ernie's actions do not have another agent as a causal source. For an agent to fail to be responsible in general, other agents must be the causal source of a *sufficient proportion* of their actions, such that it seems right to say that they *generally* lack autonomy. This is the case in the modified zygote case in which Diana designs Ernie such that her intervention is the causal source of *every* action he performs.

The second reason that I think that Agentless Sources supports ordinary humans' autonomy is that human actions and abilities have potential causal sources in evolutionary and other non-agent forces. Ernie, for example, did not evolve, he was designed by Diana, but he *is* human, and humans have the physical and biological features they do because certain physical and evolutionary forces caused them to[70]. This is why there is a meaningful difference between whether Diana designs Ernie such that he performs *every* action, or such that he only performs a single action and leaves the remaining dispositions as they would be if Ernie were naturally conceived. Even if Diana selected the remaining dispositions entirely randomly, random forces are non-agent forces and, under Agentless Sources, not autonomy-undermining (an idea supported by Herdova (2021) and Musiał (2022)).

[70] Even Diana *must* design Ernie's zygote to be a certain way for him to be able to develop in the womb, and subsequently live an ordinary kind of life. If she selects these features disinterestedly, the causal source of Ernie's dispositions is likely to be the physical and biological make-up of humans.

To summarise, assuming Agentless Sources is true, ordinary humans are autonomous *despite* the actions of other humans around them, because those actions are not causal sources of a sufficient proportion of an ordinary human's actions, and they are autonomous because their actions *have* potential causal sources in non-agent forces, such as randomness or evolutionary forces. These, along with the influence of ordinary humans frequently being the causal sources of their *own* actions, mitigate the likelihood that ordinary humans will fail to be responsible.

You might wonder whether it is appropriate to say that some human actions have causal sources in *evolutionary* or other non-agent forces. I'll briefly make the case here.

The clearest example of potential evolutionary causal sources is inherited genes. These can cause actions by determining an agent's appearance, growth patterns, psychological dispositions, etc. Many of these are realised through changes in *physiology*. One example: some people are much taller than others. Tall people's genetics led to a physiological feature, which in turn often leads to psychological or action-related dispositions (such as having beliefs about how the weather is up there or becoming a basketball player). Establishing the causal source of these actions is complex. For many actions that might be traced back to genetics there may be competing explanations that refer to childhood or upbringing. The smart money is on a mixture of both – some genetic predispositions may be causal sources of actions, but most will only be contributing causes.

Another example of the causal power of evolution is in how human physiology leads to actions more generally. Human physiological responses are the leading cause of human desires for food, water, and companionship (and corresponding actions), and they played a significant role in human's ability to use tools, our ability to communicate, to run, to breathe, etc. Without evolved physiology humans would lack these abilities and desires. It may be a struggle to see how this can a causal source of human actions since the abilities involved are so general. But consider human's *physiological* disposition to desire both sugar and salt. Humans cannot have any choice in desiring these chemicals (for good evolutionary reasons). Similarly, humans have established physiological responses to danger and newborn babies that motivate action. These desires are not a quirk of hereditary genetics, they are a consequence of evolved physiology and subsequent action-motivation neurochemistry.

To me it is highly plausible that evolutionary causal forces offer a protective effect against autonomy-undermining interventions. Humans are autonomous if evolutionary forces are the

causal sources of their actions. Often, because human reasoning and decision-making is the causal source of their own actions, evolutionary forces are not causal sources. Both conserve autonomy. But the protective effect is against another agent's interventions (and other external causes) – an intervening agent needs not only to be a stronger causal force than the humans independent reasoning, they also need to be a stronger causal force than the humans evolved motivations and dispositions. So, it seems harder to undermine an evolutionary agent's autonomy, especially when attempting to design an agent by modifying an evolutionary agent. Almost all artificial agents lack this protective effect - the designer's actions are an unchecked causal source for their actions.

Jurgen Habermas (2003) makes a similar point when arguing that genetic engineering undermines autonomy. He says that "[t]he conditions … of nature-like growth … alone allow us to conceive of ourselves as the authors of our own lives." (Habermas, 2003, 50, nt. 5); something readily interpreted as meaning that being affected by natural forces like evolutionary forces conserves autonomy. Though, I do not think evolution is necessarily the *only* autonomy-conserving external force.

A final point is that random chance is a potential causal source of human action. This might be true on several interpretations. First, non-evolutionary natural forces can be seen as a kind of randomness. Being 'randomly' hit by lightning, for example, might be a powerful causal source of a human's actions. Second, literal randomness, like a lottery. As long as the lottery is fair, human agents' actions are not the causal source of *which* human wins the lottery[71]. The force here is 'random chance', and it is sometimes a causal source of action without, according to Agentless Sources, undermining autonomy. Other aspects of random chance are discussed in theories that refer to 'luck'[72], such as where and when you are born.

To summarise the point, I have argued that Agentless Sources does not entail that ordinary humans are not autonomous. This is because ordinary human development involves local but not global autonomy undermining events, and because humans are protected from having their autonomy undermined by evolutionary forces, and to a lesser extent random chance.

---

[71] Neither the action of organising the lottery nor buying the ticket are a *causal source* of who wins the lottery. Since buying a ticket (even buying *that* ticket) does not reliably, robustly, or invariably cause winning the lottery.

[72] See Mele (2008, 2020) Cyr (2019, 2020) and Levy (2009) for discussion of manipulation arguments and moral luck.

A second objection to Agentless Sources is that some think non-agent forces can undermine autonomy. Watson and Frankfurt suggest this:

> "It is hard to see what differences there could be between the natural and purposeful forms of determination that would be relevant to freedom and control" (Watson, 1999, p, 67).
>
> It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberately manipulative designs of other human agents. (Frankfurt, 2002, p. 28)

Fischer rejects agent-based historical conditions for similar reasons as those he used to reject intentional conditions. He says:

> Might it be that there is an agent involved at the beginning of the sequence—an agent who creates Ernie in just the way necessary and sufficient for Ernie's subsequent behavior (doing A)? Perhaps Diana accidentally creates Ernie in just this way, without intending that Ernie will subsequently do A. This is certainly imaginable, but it is completely obscure how this fact—the fact that there is an agent involved, an agent without any relevant intentions—should make a difference as to Ernie's subsequent responsibility status. (Fischer, 2016)

Mele (1995, p. 168) suggests that there is no difference between being manipulated by an agent and passing through an electrical field with the same effect. Suppose that this happens to Beth with the result that she becomes psychologically identical to Ann as in the Ann/Beth case. Call her 'Bermuda Beth'. Mele claims that Bermuda Beth has her autonomy undermined[73]. However, Agentless Sources implies that Bermuda Beth *does* meet the history condition for autonomy. To the extent that this is unintuitive, I am happy to bite the bullet here. But it does not seem unintuitive to me, after all, while electrical fields are not a common cause of human dispositions, *other* non-agent forces are consistent with autonomy. The forces that lead to 'morally lucky' situations, such as being born in Nazi Germany, or being, by chance (or genetics), entirely incompetent at performing evil acts do not seem to undermine autonomy or responsibility. But an agent arranging equivalent situations, such that they are the causal source of your actions in those contexts, does seem to be autonomy-undermining and responsibility relieving. Similarly, if an *agent* arranged for Bermuda Beth's electrical

[73] Pereboom (2001) presents a similar case where machines intervene with Plum in the four case argument. These cases are discussed by Liu (2022) and De Marco (2023) – but Bermuda Beth suffices here, and my answer applies to the machine case too.

storm, then her autonomy *would* be undermined because that agent would be the causal source of her actions. To me, at least, this seems to track standard autonomy and responsibility ascription practice. So, I see no reason to think that non-agent forces can be autonomy-undermining causal sources.

Two final clarifications. First, it is important that we distinguish between times when an agent's action is the causal source of another agent's actions and times when an agent's action is causally necessary but not a causal source for another agent's actions. For example, when conceiving a child, parents' actions are not the causal source of the future child's genetic composition. Parents actions are merely a conduit for the random, cellular, and evolutionary processes that determine the child's genetic make-up. Choosing to conceive does not have a strong enough causal relationship with the child's actions to be a causal source, and thus, (as we would hope and expect) is not autonomy-undermining according to Agentless Sources. But choosing to design an agent is often the causal source of a designed agent's actions, because there is often no alternative causal source (such as evolutionary forces).

Second, Agentless Sources applies to human and non-human systems indiscriminately. Animals, for example, are normally autonomous because evolutionary/physiological drives are the causal source of most of their actions. Thus, the Agentless Sources history condition of autonomy does not privilege humans or even those who obviously possess (internal) intentional states.

In summary, this section has laid out Agentless Sources as a historical condition of autonomy that can satisfactorily distinguish between ordinary humans and designed/manipulated agents. It thus enables denial of the no difference claim and facilitates a soft-line reply to design cases. There is, according to Agentless Sources, a responsibility relevant difference between designed/manipulated and responsible agents. In the following sections I will lay out the ramifications, which are that the Design Hypothesis false but a nearby hypothesis is true: the act of design undermines autonomy unless the designed agent can perform actions with non-agent external causal sources.

7.8. The Design Hypothesis Reconsidered

A reminder of the Design Hypothesis.

Design Hypothesis: The act of design always causes a designed agent to fail the conditions for autonomy needed to be responsible.

If Agentless Sources is true, as I have argued, then The Design Hypothesis is false. However, a nearby hypothesis is true. Acts of design are normally the causal source of a majority of designed agents' actions. By defining a function, and creating or modifying material to reflect that function, acts of design are likely to have the strongest causal relationship with a designed agent's action. If they were to design the agent with a different function in mind, the agent would act differently, if the background conditions were different, the designers would change the designed agents functioning such that it achieved the same purpose. (For example, designing a golf ball on the moon would require accommodating for the decreased gravity by using a material with greater mass). Thus, in most cases acts of design are autonomy-undermining, and the designed agents involved fail the historical condition of autonomy and cannot be responsible. This is true for all designed systems that exist presently, and for all artificial systems being produced through traditional machine learning methods (probably including ones that are nominally evolutionary, since their 'evolutionary' development will have lower causal variance than the designer's actions). So it is true that *many* acts of design are autonomy-undermining and the resulting agents cannot be responsible.

However, the Design Hypothesis, if Agentless Sources is true, is false because it is possible to perform the act of design randomly or guided by natural forces, thus preserving the designed agents' autonomy. Musiał (2022) discusses random design of agents as a potential means of designing agents non-manipulatively. In this, we agree: random forces are not autonomy-undermining. However, Musiał is much more optimistic about the possibility of randomly generated agents than I am. I am pessimistic because the level of randomness may be a problem. Musiał suggests that a randomly generated agent only requires that their desires are randomly chosen; however, in this case it seems to me that a designer's creation and modification of the list of desires has a sufficiently strong causal relationship to the 'randomised' agents actions to undermine autonomy. Randomly generating the desires themselves would probably weaken the causal relationship to the designer sufficiently for the random forces to be the causal source of the designed 'random' agents actions, but I suspect that no coherent agent can be produced in this manner.

Another proposal worth considering is to intentionally create an autonomous agent, such that the designer's actions do not have a strong causal relationship with the designed agent's actions and are therefore not a causal source of the designed agents actions. However, while a designer may *intend* to create autonomous agents, I am not sure how they would *succeed* to do so (unless adopting my evolutionary method below). A designer may confer the proper function of 'being autonomous', 'having no function' or 'being able to decide their own function' and create an agent intended to be so; but to succeed in this they would need to modify or create the designed agent's material constitution such that they fulfil that function. In doing so they would confer lower-order functions: give the designed agent a brain so that it may think for itself, give it eyes so that it may see for itself, etc. But the act of designing these constituent parts seems to be a fresh route to being a causal source – the designed agent finds their actions limited by the actions of the designer, with their physical system, they can only perform certain types of actions in a certain way. The causal source of these actions, despite the noble intentions of the designer, will still be the act of design, and the resulting agent will thus still lack autonomy. You might think that the abstract nature of the overall function of the designed agent would be sufficient to make it the causal source of its own actions, but I disagree – it seems to me that all the details of the designed agent's individual actions can be traced back to the act of design. Imagine a designed agent, for example, that cannot fly, that cannot breathe, that lacks phenomenal consciousness, or creativity, cannot burrow under the ground or survive underwater – all these qualities, should humans have them, could be had autonomously if they were the product of evolutionary forces, but in this case they are the direct consequence of a designers (perhaps unintentional) actions, and thus are autonomy-undermining. Autonomy cannot be grounded in noble intentions, and to be a designer of an agent with the function of autonomy is to be Daedalus giving Icarus waxen wings.

To produce an autonomous, and thus possibly responsible artificial agent, a different methodology is needed. I suggest that this methodology needs to be integrated with the primary non-agential causal source of human action: evolutionary forces. Agents that evolve can track the causal source of their actions to evolutionary forces, which offers a protective effect against having their autonomy undermined compared to straightforwardly designed agents. Therefore, a designer who designs an evolutionary system may be able to avoid being an autonomy-undermining causal source of the designed agent's actions. Even so, as I discuss in the final chapter, evolved artificial systems face several potentially manipulative scenarios,

and the design of evolved artificial agents (artificial life) must be done carefully to avoid undermining the autonomy of those agents through design choices as in the previous paragraph.

Part III: The Evolutionary Pathway to Artificial Moral Agency

8. The Evolution of Human Moral Reasons-Responsiveness

8.1. Introduction

In chapter 6 I argued that artificial systems *can* be morally reasons-responsive. However, to do so, they must be reliant on deference, which is seen as a flaw. At this point, it can be seen that one good reason why deference undermines moral agency is because the act of deference undermines autonomy by shifting the causal source of moral actions outside the agent.[74]

In chapter 7 I argued that artificial systems will not meet the historical condition for autonomy, and therefore will not be responsible, unless they are designed in a way that shifts the causal source of the designed agent's actions to natural or random forces. I have suggested that evolutionary dynamics and processes are a good choice for this. In chapter 9 I will argue that evolved artificial agents *can* be autonomous, and furthermore, evolved artificial agents can develop adequate moral reasons-responsiveness without having their autonomy undermined – unlike deferential agents or most designed agents.

This chapter establishes the groundwork for that argument. It describes how humans evolved to be moral agents. To do so, it draws on theories about 'the evolution of morality', sometimes called 'evolutionary ethics'[75]. While it is uncontroversial that humans evolved to be moral agents, theories about the evolution of morality adopt different approaches and emphasise different aspects of that evolution[76]. Here, rather than settle any scores between them, I broadly identify the evolved human capacities that are sufficient for moral agency.

8.2 recapitulates the concept of moral agency and clarifies the type of evolutionary explanation being offered. 8.3, 8.4 and 8.5 divide the evolved capacities for moral reasons-responsiveness into three: *biological* capacities, *psychological* capacities, and *conceptual* capacities. Humans are adequately moral reasons-responsive using a mixture of all three capacities. While it may be possible to be adequately moral reasons-responsive in other ways,

[74] Though that does not likewise explain why *moral* deference is *especially* problematic.
[75] Which is the successor to 'sociobiology' (Wilson, 1975).
[76] Influential works include (Alexander, 1987; Joyce, 2006; Kitcher, 2011; Tomasello, 2016; Waal, 1997)

I argue that a designer aiming to make artificial moral agents is best served by aiming to design artificial agents that will evolve human-like capacities.

Conceptual capacities are the most complex and the least common in the natural world, so they are likely to be the biggest challenge for an evolutionary artificial agent to develop. To aid in facing this challenge, 8.6 presents some key conditions for the evolution of conceptual capacities in protohumans. If these key conditions can be recreated for an evolving artificial agent, then the chances of successfully evolving an artificial agent with conceptual capacities may improve.

8.2. The Evolution of Morality

The concept of moral agency I have been using has two conditions: to be a moral agent an agent must have the capacity to adequately respond to moral reasons and the capacity to be responsible.

The key issue of the previous chapter was the historical condition of autonomy, which must be met for an agent to be responsible. Systems that evolved through natural selection have the right kind of history to be autonomous – their actions' causal sources can be evolutionary forces (or themselves), just like humans' actions can. So, I will assume that evolved agents can satisfy the historical condition of autonomy for responsibility (though in the following chapter I will discuss some potential issues with designing artificial agents to evolve without undermining their autonomy). Of course, there are *other* conditions of responsibility: responsibility has both epistemic and nonhistorical rationality or (general, rather than specifically moral) reasons-responsiveness conditions. This chapter centrally focuses on the evolution of capacities for moral reasons-responsiveness, but in the conclusion, I suggest that the same capacities can also perform the necessary roles for satisfying the nonhistorical and epistemic conditions for responsibility. That is, an agent that evolves adequate moral reasons-responsiveness without an autonomy-undermining history is likely to be satisfy the other conditions of responsibility too, and therefore likely to be a moral agent.

Competing explanations for the evolution of moral reasons-responsiveness emphasise different processes, causes, and capacities. In this chapter I attempt to provide a general explanation that can be used as a guideline or template for evolving artificial moral agents.

I now offer some preliminary clarifications. First, I will discuss descriptive, rather than prescriptive accounts of the evolution of moral reasons-responsiveness. Second, given my understanding of moral reasons (restated below), I assume that simple behavioural dispositions can provide a crude responsiveness to moral reasons. Third, the chapter's focus is on the evolution of the capacities for adequate moral reasons-responsiveness. I do not discuss which developments were *necessary* for this, them being jointly sufficient is enough. I will now explain these points further.

I use a 'descriptive' account of evolutionary ethics here. 'Prescriptive' accounts of evolutionary ethics claim that evolutionary forces somehow determine what is morally right or wrong (FitzPatrick, 2021). They are not the target theories here. That something evolved or is likely to evolve is not taken to determine if it is *right* or *wrong*. A parent's inclination to abandon their child in grave circumstances might be likely to evolve if it leads to greater 'fitness' (i.e., chances of reproductive success). But that does not, I assume, make it *right*. 'Prescriptive' evolutionary ethics is controversial, and its truth is unnecessary for my arguments. As usual, I aspire to avoid metaethical issues too: perhaps the evolution of moral reasons-responsiveness bears consequences for the best theory of moral facts[77], but even if so, that is tangential to my claims.

The theories I am about to discuss are instead 'descriptive' accounts of evolutionary ethics. The descriptive claim is that human's capacity to respond to moral reasons evolved, like many of our other capacities did. 'Descriptive' theories look to evolutionary history and evolutionary forces to explain *how* this happened, what was instrumental in this occurrence, and how understanding the evolutionary origins of morality might improve our moral theories and understanding.

William Fitzpatrick provides a modest, and relatively uncontroversial, summary of the descriptive claim:

> *The Modest Evolutionary Explanatory Thesis*: evolutionary forces may adequately explain certain *capacities* and *tendencies* associated with moral thinking, feeling and behavior, and may explain or *partially* explain *some* of the *content* of our moral thought, feeling and behavior, insofar as it is influenced (individually or via influences on cultural development) by those tendencies. (FitzPatrick, 2021)

[77] As, for example, 'evolutionary debunking arguments' (Street, 2006) argue.

There have been efforts made to defend this thesis. Indeed, accounts of the evolution of morality often consider the vindication of this explanatory thesis to be a core aim. However, I do not defend it here, but instead assume its truth. It is reasonable, I suggest, to assume that there is an evolutionary explanation for the human capacity to respond to moral reasons. Mirroring Fitzpatrick's qualifications, this does not mean there is a full and complete evolutionary explanation of all moral behaviour or the content of all moral beliefs. My assumption is just that humans' general capacity to respond to moral reasons has an evolutionary explanation.

Having adopted the descriptive account. The second clarification: given my understanding of moral reasons, even simple evolved capacities can exhibit moral reasons-responsiveness. It is safe to assume that humans respond to moral reasons. However, some think other animals cannot respond to moral reasons *at all*. As discussed in 3.4., I understand moral reasons to be a domain of reasons that seem to be centrally described by moral concepts and that plausibly play a role in a moral theory.

To be clear, on my understanding of moral reasons, simple animals (like insects) can respond to moral reasons. They cannot necessarily *adequately* respond to moral reasons (i.e., they cannot necessarily meet the moral reasons-responsiveness condition of moral agency). Rowlands argues for a similar position in detail – for him, animals have some degree of moral reasons-responsiveness, but this capacity is inadequate for moral agency.

Here it is worth explaining that moral reasons-responsiveness as I conceive it operates along two axes. The moral reasons involved have a *range*, such that an agent which responds to a *greater range* of moral reasons is more morally reasons-responsive than an agent which responds to a lesser range. And the reasons involve are *vague* or *precise*, such that an agent which responds to precise moral reasons is more morally reasons-responsive than an agent which responds to vague moral reasons. Precision increases the *volume* of moral reasons responded to – for example, if I precisely respond to the interest of every individual in a country, I may respond to 50 million moral reasons, whereas if I vaguely respond to the general interest of the country's inhabitants, I may only respond to one reason. But greater precision does not increase the *range* of moral reasons – my 50 million reasons are all more or less the same as each other. *Range* tracks the *type* of moral reason – the set of moral reasons about all sorts of agents in all sorts of situations has a greater range than the set of moral reasons about one agent in one situation. *Adequate* moral reasons-responsiveness is the

capacity to respond to a set of reasons that are equivalent (or greater) in both range and precision to the set of reasons that the average human responds to.

But, as mentioned, some think that animals cannot respond to moral reasons at all. One argument for this can be taken off the table immediately. It is true that if only *adequate* moral reasons-responsiveness counts as genuine moral reasons-responsiveness, then of course most (and perhaps all) animals do not respond to moral reasons at all[78]. But this is a verbal rather than substantial issue – the definitions of the domain of moral reasons differ.

Assuming my understanding of moral reasons, you might still want to push the case for animals being unable respond to reasons that are centrally moral and that can be plausibly used in a moral theory. I disagree (and I will offer reasons why throughout the chapter) but even if this were true it would not be terminal for my position. Suppose, for whatever reason, that many animals do not respond to moral reasons at all. Their capacities may still be relevant to the evolutionary development of moral agency if they contain some influential progress towards achieving adequate moral reasons-responsiveness. As will become apparent, almost all theories of the evolution of morality claim that their capacities *do* offer some progress of this type. I think that they also exhibit some degree of moral reasons-responsiveness, but even if they do not, they still play a role in the developmental template I am looking for.

A related point concerns whether any or all these capacities are *necessary* for the evolution of adequate moral reasons-responsiveness. They are necessary in this sense if any agent without these capacities at some point in their evolutionary history could not evolve moral reasons-responsiveness. The aim of the chapter is to present a template, or a pathway, for the evolutionary development of adequate moral reasons-responsiveness – not to identify its necessary conditions. Essentially, the evolution of morality is a story that concludes with adequately moral reasons-responsive agents and any agent that follows the same story should satisfy any necessary conditions, should they exist.

The next sections outline the types of evolved capacity that facilitate moral reasons-responsiveness. Most descriptive theories of the evolution of morality offer detailed accounts

[78] We might understand Korsgaard (2006) as adopting this kind of definition in her objection to De Waal's claims about ape morality (Waal, 2009)

of various evolved moral capacities (Alexander, 1987; Bekoff & Pierce, 2010; Curry, 2016; Gibbard, 1992; Graham et al., 2011; Rottschaefer, 1998; Rowlands, 2012; Sober & Wilson, 1998; Sterelny, 2014; Tomasello, 2016; Waal, 1997; Waal et al., 2014). My discussions over the following sections is centrally informed by Richard Joyce (2006) and Philip Kitcher's (2011) excellent books on the evolution of morality – their accounts are two of the most developed in this area. The capacities can be summarised as the following. Biological capacities are biological impulses to act in a way that corresponds with crude moral reasons, such as acting to help another. The standard example of an animal with biological capacities (but not much else) are bees. Psychological capacities confer the ability to respond to moral reasons through psychological skills such as the representative of joint aims, emotional reactions to the intentions and behaviour of others, and advanced social problem-solving. The standard example of an animal with psychological capacities like these are chimpanzees. Conceptual capacities enable moral reasons-responsiveness through abstract moral concepts and moral understanding. On Earth, (probably) humans alone have conceptual capacities.

All these capacities, as I will discuss, are united in generating behaviour that responds to moral reasons, but their mechanisms, and correspondingly the range and precision of the moral reasons they enable responsiveness to, differ. Biological capacities are not adequately moral reasons-responsive – the range and precision of the set of moral reasons being responded to is small. Psychological capacities respond to a greater range of more precise moral reasons, but still fall short of adequacy. Conceptual capacities respond to the greatest range of most precise moral reasons; but they do so by refining the 'raw material' of psychological capacities.

8.3. Bees

The simplest type of evolved capacity to be discussed is simple biological dispositions to help others. One popular example is social insects like bees. Bees are both social and cooperative (Joyce, 2006, Chapter 1). They work together without harming each other, they each contribute to the working of the hive, they collect and distribute resources together, sometimes they police each other's behaviour (Ratnieks, 1988; Ratnieks & Wenseleers, 2008), and they will sacrifice themselves by stinging to defend their hive. They clearly lack advanced moral sensibilities – a bee does not reliably respond to a variety of moral wrongs like a human would.

The standard example of cooperative, or 'prosocial', biological behaviour is 'fitness sacrificing' behaviour (Joyce, 2006; Kitcher, 2011). Fitness sacrificing behaviour is behaviour which benefits others at a cost to actor[79] . Individual bees are fitness sacrificing in a few obvious ways: first, most worker bees are sterile, innately sacrificing their chance at reproduction but helping the hive in general; second, bees respond to threats by sacrificing themselves; third, bees contribute to collective efforts by depositing and sharing resources rather than taking them all for themselves.

'Fitness sacrificing' behaviour superficially goes against evolutionary principles. The helping bee reduces their all-important chance of reproducing. The evolutionary explanation for fitness sacrificing behaviour is that while the bee decreases its reproductive chance as an individual, its action increases the reproductive chances of the other bees by an even greater amount. Since the bees are genetically related, then the overall reproductive chances of the genes increase – this is what is called 'kin selection' (Hamilton, 1964; Queller, 1992; Queller & Strassmann, 1998; Trivers, 1971). There are similar explanations for other simple animal behaviour which involves working together: cooperating often yields a reproductive advantage. For example, mutualisms (such as wolves hunting in packs) and parental care clearly offer a reproductive advantage and thus have an easy-to-access evolutionary explanation.

While bees' fitness sacrificing behaviour has an evolutionary explanation[80], it still seems to respond to a moral reason. I will use the rough guideline of whether it is centrally described by moral concepts and can be plausibly used in a moral theory to demonstrate. First, the bee sacrifices its own interest in the interest of others, and this is readily described by moral concepts. After all, a human sacrificing their own interest for another is described in moral terms. You might morally praise a human that did so, saying, 'she nobly overcome her selfish desires and spent her efforts to contribute to the greater good'. In 'A Christmas Carol' Scrooge transforms from *morally reprehensible* to *morally praiseworthy* because he transformed from self-interested to self-sacrificing. From a moral reasons perspective, these humans are not responding to moral reasons *because* they overcame their self-interest, they

[79] Many sociobiologists call this behaviour 'altruistic'. Richard Joyce (2007) dislikes this terminology and I prefer not to use it for similar reasons.
[80] Though the effect of kin selection can be overstated, and there are some concerns about how well it explains social insects like bees behaviour (Nowak et al., 2017)

just began to respond to a moral reason that they did not before. The bee is the opposite – its biological dispositions cause it to *always* respond to moral reasons about sacrificing itself for the greater good of the hive. Which does not lessen the fact that it responds to this moral reason. One way of expressing this moral reason is that, generally speaking, you should help others who are in need, even when it goes against your own interest. Most human moral theories do indeed hold such a reason to be true, which is the best possible evidence for its plausibility as part of a moral theory. So, since bees biologically driven cooperative and fitness-sacrificing behaviour is responds to a reason that is centrally described in moral terms and plausibly part of a moral theory, then it responds to a moral reason. It may not do so *precisely*, and it may not respond to a great range of moral reasons, but, nonetheless, its biological capacities exhibit responsiveness to moral reasons.

The same is true for other types of biologically driven cooperative behaviour. Mutualisms promote the needs of others as well as oneself and parental care can be understood as protection of the vulnerable or an improvement in wellbeing (Joyce, 2006; Rottschaefer, 1998, Chapter 2; Rowlands, 2012). These kinds of actions are moral reasons-responsive as I understand it – even though the range and precision of those moral reasons are low. I call this group of capacities the 'biological capacities' for moral reasons-responsiveness – generally just 'biological capacities', for short.

Some argue that all capacities for moral reasons-responsiveness stem from evolutionary forces like kin selection – Joyce (2006) says, "the unquestionable importance of the mother-child relation is sufficient for us to conclude without going to much trouble that kin selection was an important force in our heritage." Oliver Scott Curry (2016), Kitcher (2011), and Michael Tomasello (2016) all broadly support a similar claim: the evolution of biological capacities in animals was a central plot point in the development of human morality. Humans, they think, would never have become the moral reasons-responsive creatures they are today without their evolutionary heritage of biological capacities.

Of course, bee and human moral reasons-responsiveness is vastly different. Humans respond to moral reasons by conceptually understanding them and deliberating about them, while bees do so through unreflective, rigid, and uncontrollable biological capacities. Though that is not to say that biological capacities for moral reasons-responsiveness are absent in humans. As Joyce intimated, humans have biological capacities of their own. Empathic neurons lead them to involuntarily wince when seeing another human in pain. They cry out in pain, signalling

danger to others rather than staying hidden. Exposure to babies prompts a flood of hormones that produce a strong caring impulse. The act of breastfeeding lowers the fitness of the mother to increase the fitness of the child. All these seem largely driven by involuntary biology rather than voluntary psychology (See Churchland, 2021 among others).

There remains a serious difference between human and bee behaviour, even when human biological capacities are in play. Human hormones and other sources of caring actions do not themselves produce action. Normally a psychological mechanism is also involved. Parents do not act to help their child based on uncontrollable biological impulses like bees do, but instead human biochemistry prompts an *emotion*, or *motivation* to act (See especially Kitcher, 2011, Chapter 2 on what he calls 'psychological altruism'). This motivation must still pass through a process of psychological deliberation to result in action.

A human with a biological drive to care for their child can suppress those impulses if they believe that acting on them would be harmful. This is one difference between human moral reasons-responsiveness and bees' moral reasons-responsiveness. Adequately responding to moral reasons involves the ability respond to a range of moral reasons. Biological capacities can do this in principle, a bee might evolve the biological capacity to avoid self-sacrificing in certain situations, and that set of situations might eventually track a large range of moral reasons. But in practice humans use psychological and conceptual representation to expand their range of moral reasons – they suppress their biological drives at times. Bees, after all, cannot respond to a wide range of moral reasons– they cannot adapt to novel moral situations, and they cannot reliably respond to moral reasons in complex situations. The range of moral reasons they respond to is limited to a few moral reasons about the value of the interests of others in specific contexts. So, bees, and other species on Earth with biological capacities (but no others) for moral reasons-responsiveness are not *adequately* morally reasons-responsive.

Biological capacities form part of the human mechanism for moral reasons-responsiveness that *is* adequate. However, sometimes humans respond to moral reasons based on their *psychological* or *conceptual* representation of that reason without any obvious or immediate corresponding biological capacity. For example, I might believe that I should donate to charity based on conceptual moral reasoning alone. My acceptance of these conceptual moral reasons may be somehow dependent on my having biological capacities at one point, but at the point of performance my action of donating may not involve them. So, biological

capacities form part of human, adequate, moral reasons-responsiveness, but humans do not use biological capacities every time they respond to a moral reason.

In summary, bee-like biological capacities are not adequately moral reasons-responsive. They may, however, be a valuable part of the development and performance of human moral reasons-responsiveness. As I will argue in the following chapter, developing capacities like these in artificial agents is easy enough. However, doing so means little without developing capacities for greater moral reasons-responsiveness.

8.4. Chimpanzees

On Earth, biological capacities alone have been insufficient for adequate moral reasons-responsiveness[81]. Other capacities play an important role in human moral reasons-responsiveness that goes uncompensated for in bee-like biological capacities. Some animals, typically big-brained highly social mammals, have what I call 'psychological' capacities for moral reasons-responsiveness. These involve things like emotions, intuitions, and representations. They seemingly respond to a greater range of more precise moral reasons than biological capacities.

Chimpanzees are the standard example here (in part because of Frans De Waal's influential work on chimp morality Waal, 1997). As Joyce says: "The mechanisms in place that determine the helping behaviors of bees are unlikely to bear much resemblance to those that ensure the helping behaviors of chimpanzees. The evolutionary processes that explain such helpful behaviors may be broadly the same (it may be kin selection in both cases, for example), but the means by which those processes achieve results are going to differ remarkably." (2006, p. 44). There are three primary features that improve chimps' capacities for moral reasons-responsiveness compared to bees. First, chimps can respond to moral reasons emotionally. Second, they can represent others' intentions and goals. Third, they can have conflicting desires and can overcome self-interested desires to benefit others.

Before explaining these three capacities in more detail, I will explain a little more of the evolutionary context. Chimps, unlike bees or (mere) group hunters, behave cooperatively in

[81] Though perhaps they are *possibly sufficient*, they have not been sufficient on Earth.

ways that promote fitness via what is known as 'indirect reciprocity' (See Alexander, 1987, Chapter 2 for an early account of evolutionary ethics focusing on indirect reciprocity). This is, essentially, a reputation-based system. Chimps do not (just) cooperate with one another on a directly reciprocal 'I help you because you help me' basis; they cooperate with one another within a changing group dynamic (Kitcher, 2011; Silk, 2007; Tomasello, 2000; Waal, 1997, 2006). Chimps form unique relationships with other members of their social hierarchy. There is a leadership group, containing the leader and their lieutenants; and other members form and dissolve alliances and coalitions depending on the group's situation (the group could be, for example, facing leadership change, recovering from hardship, or rearing young). Leaders come to power through both physical might and social cunning, seeking beneficial alliances (Melis et al., 2006) that strengthen their hold on power. In all this, reputation is central. If able, chimps tend to reward their helpers and punish their opponents, but a single helping action isn't enough to turn the tide, chimps tend to help their trusted allies more than capricious chimps who just happened to be on their side (Gilby & Machanda, 2022; Schmidt & Tomasello, 2016; Suchak et al., 2016; Suchak & Waal, 2016).

This social organisation promotes fitness individually, as individual members benefit from the division of labour in hunting, childcare, and defence, and via kin selection, since the entire group benefits and group members are often related to one another. But to reap those benefits and succeed as a group, chimps require advanced social and cognitive skills. It is in this context that the evolution of greater capacities for moral reasons-responsiveness can be explained. Since closer social organisation offered advantages in fitness, genes which correlated with better capacities to respond to moral reasons tended to persist. It is likely that mammals were best placed to evolve these capacities. Rowlands surveys evidence for animal moral reasons-responsiveness in the form of 'moral emotion' and suggests that it is in "social mammals— elephants, gorillas, chimpanzees, monkeys, dogs, and rats … that the case for possession of moral emotions is strongest." (Rowlands, 2012, p. 46).

Having discussed why psychological capacities for moral reasons-responsiveness evolved, I turn to features of chimp's psychological capacities. First, chimps respond to some moral reasons emotionally. They respond angrily to cheating, sympathetically to injuries, caringly to loyalty, contritely when they have harmed another, etc. Some debate whether these actions

truly 'emotional'[82] (Rowlands, 2012). But, whatever their proper name, chimps' respond to moral reasons in this manner. Chimp moral reasons-responsiveness is distinct from bee-like moral reasons-responsiveness because chimps have a mediating 'layer' of emotional representation which enables them to respond more flexibly and more precisely to moral reasons.

One interpretation of chimp moral-reasons responsiveness is that it is no more than a complex biological capacity for moral reasons-responsiveness – they are just upgraded bees. But the charitable interpretation (Championed by de Waal, 1996 and elsewhere) is that they are driven by genuinely other-regarding moral motivations, albeit ones that they do not always control or understand. This latter interpretation has a lot going for it. After all, humans motivated by feelings of concern or sympathy act in a genuinely other-regarding manner.

Both chimps and bees can act against their own interest to benefit others. But chimps can respond to more precise moral reasons because their emotions are context-sensitive and track many features in a situation (See Godfrey-Smith, 2018; especially Sterelny, 2003; Tomasello, 2016). Chimp's emotional behaviour respond to moral reasons more precisely because different moral reasons prompt different emotions. Rather than responding to a single, vague, moral reason in a situation, chimps can respond to the multiple, more precise moral reasons through varying intensities and combinations of emotions. (Kitcher, 2011, Chapter 1).

Not only is chimp reason-responsiveness *more precise*, but it is also *wider ranging*. Chimps can emotionally respond to moral reasons about resource distribution, like promoting fairness and equality, and about harm, like when punishment and guilt are appropriate. Because of the highly context dependent nature of these moral reasons, which depend on histories of individuals and specific circumstances; they are unavailable to animals like bees that only use biological capacities. Emotions allow chimps to respond to a greater range of moral reasons and to respond to moral reasons more precisely.

In responding to moral reasons, chimps pair emotions with the second capacity: the capacity to respond to the intentions, aims and desires of others (Argued for in detail by Tomasello, 2016). Chimps cooperate with one another to achieve aims, including group aims. They have

[82] One tactic is to call them 'sentiments', which tends to be a successful appeasement.

complex relationships with one another that imply a rudimentary understanding that others are agents, with needs and desires of their own. Chimps can act to promote or frustrate other chimp's interests through internal, psychological, representation of those interests. Compared to bees, which also cooperate and live in groups, chimp groups contain complex and changing social relationships, they can work together to respond to changes in the environment – exploiting opportunities that arise from the actions of other chimp groups, the ripening of fruit, and changes in prey activity or weather. These behaviours are evidence chimps' have the basic psychological capacity to represent other minds. Though they do not have a *full* capacity to respond to others ongoing evaluations (Engelmann et al., 2012; Engelmann & Tomasello, 2018). They also represent a more precise responsiveness to moral reasons – chimps have the equipment to respond to others interests qua interests (and reasons for promoting others' interests are *moral reasons*[83]). Acting on group aims responds to moral reasons about a *group's* interest[84] and thus increases the range of moral reasons chimps respond to compared to bees.

Finally, chimps have a third capacity: the capacity to mediate between conflicting desires. They can be uncertain about how to act in a moral situation. Consider the following example from Philip Kitcher:

> Chimpanzees are openly torn between selfish and altruistic courses of action, making it apt to attribute to them two desires, both expressed in facets of their behavior. An animal hesitates. Holding a branch rich in leaves, he is poised to strip them off and eat, and, simultaneously, the set of the body acknowledges the presence of an ally; eventually, the arm is extended, thrusting a small bunch of leaves toward the friend, while the rigidity of the gesture and the averted face show the presence of a contrary desire. (Kitcher, 2011, p, 72)

Being uncertain like this is indicated by chimps' *sometimes* performing helping actions and other times satisfying their own desires in similar circumstances. That chimps do this implies a nascent capacity for what Gibbard (1992) calls 'normative guidance' – which is the capacity to make principled decisions between competing selfish and moral behaviour. A chimp with a moral emotion is motivated, through something like sympathy, to act in

[83] Again, we can infer this because they are central to many human moral theories.
[84] Something we (but not chimps) might understand as the sum of the collective individual interests that stand to be promoted by the achievement of the group aim.

another's interest; with a rudimentary theory of mind, they can represent this interest as belonging to the other individual; with something like normative guidance, they have the beginnings of the recognition that there ought to be some kind of *solution* – i.e., a moral reason -- to the conflict between self and other-interested action. (See Kitcher, 2011, Chapter 2)

That said, normative guidance may not always appear in a *moral* context. It occurs whenever an agent chooses between acting self-interestedly and acting for *some other reason*. Kitcher offers the example of a chimp who refuses to act self-interestedly because they are afraid of others' punitive actions (Kitcher, 2011, p. 79). Again, the difference between chimp and humans should be emphasised: chimps do not reliably adhere to these non-selfish reasons, or show cognisance of moral reasons, as humans do. But chimps' ability to sometimes act for non-selfish reasons indicates the beginnings of the ability to do so.

I lump these capacities together as *psychological* capacities – as they are to be contrasted with *biological capacities*. Biological capacities result in rigid, uncontrollable, and unreflective moral behaviour; the psychological capacities just described result in flexible, controllable, and context-sensitive moral behaviour. Chimp behaviour, it seems, is much more moral reasons-responsive than bee behaviour. They achieve this through psychological capacities for moral emotions, a theory of mind, and a primitive form of normative guidance.

But is chimps' moral reasons-responsiveness *adequate*? There is disagreement here. For those who hold a low adequacy threshold for moral reasons-responsiveness, it is (DeGrazia, 1996; Sapontzis, 1992; Shapiro, 2006). But the majority think that it is not. My guideline for adequate moral reasons-responsiveness is *human equivalence*. So, I agree with the majority that an agent needs more advanced capacities than chimps to be adequately moral reasons-responsive because chimp-level moral reasons-responsiveness is clearly not human-equivalent.

Biological capacities probably play a small but significant role in the everyday performance of humans' moral actions. Psychological capacities play a starring role. At least for human moral-reason responsiveness, psychological capacities are core capacities. This is reflected in their complexity. Biological capacities are relatively easy to simulate, but psychological capacities require complex capacities and cognition which are much harder to simulate. A fact which suggests that, despite the majority opinion that chimp-level moral reasons-

responsiveness remains inadequate for moral agency, evolving an artificial system with chimp-level capacities would be a major success for the machine ethicist.

8.5. Humans

Humans, unlike chimps, are adequately moral reasons-responsive, and one explanation for this is that humans, unlike chimps, can understand, use, and represent (moral) concepts with language (Joyce, 2006, pp. 91–92; Sinnott-Armstrong & Miller, 2007). I call these capacities 'conceptual capacities'. Conceptual capacities, it has been argued, offer a significant step forward in moral reasons-responsiveness. In this section, I outline two ways in which conceptual capacities lead to greater moral reasons-responsiveness than psychological capacities. These are that conceptual capacities lead an agent to possess conceptual moral intuitions, which increase the range of moral reasons an agent can respond to, and rich moral concepts, which increase the precision of moral reasons-responsiveness. Based on this, I suggest that the adequate moral reasons-responsiveness humans possess does require conceptual capacities.

Humans, and presumably every competent conceptual language user, can have conceptual moral intuitions: the intuitive grasping of abstract, second-order moral reasons. For example, suppose the following statements are true. It is good to promote wellbeing. Having interests that can be frustrated is sufficient for being capable of wellbeing. Many animals (and perhaps even artificial agents) have interests that can be frustrated. With a conceptual understanding of these statements, humans can infer, via knowledge of logical and conceptual relations, that humans have a (pro tanto) duty to act to increase many animals' wellbeing. Acting on this belief is responding to a moral reason that can only be divined with conceptual understanding (Chudnoff, 2016; Huemer, 2005; Pust, 2016). Chimps, in contrast, are unable to believe that they have this duty, because they lack the necessary conceptual understanding of the statements and their relationships with one another. Hence, conceptual capacities allow humans to respond to a greater *range* of moral reasons.

However, those who argue that conceptual capacities are why humans are adequately moral reasons-responsive tend to focus on moral concepts rather than conceptual intuitions. Moral concepts enhance the precision and range of moral reasons-responsiveness bestowed by the

psychological capacities for moral emotions, theory of mind, and normative guidance. I will discuss conceptual capacities relationships with each of these in turn.

Humans conceptualise their emotions. By conceptualising their emotions, they can identify precise differences between similar seeming actions. They can identify the difference between emotions that motivate self-interested and other regarding action. For example, they can distinguish between being motivated by fear and by guilt, or between an angry reaction to the harm of others and an angry reaction to harm to oneself. Furthermore, they can distinguish between social emotions, such as romantic jealousy, from more primitive 'basic' emotions, such as fear or hunger.

Alone, conceptualising emotions may not significantly increase moral reasons-responsiveness. It can even lead to further complications, as emotional concepts can underpin wrongful associations between emotional experiences and stimuli. Many humans feel guilty without committing wrongs; or are fearful in the absence of threats. In many cases, this is more than mere Pavlovian conditioning. It is a deeper confusion that results from conceptualising emotions.[85] For example, many people will associate their conceptualisation of an emotion with sensations. They believe that when they feel those sensations, then they have that emotion. Furthermore, they believe that when they have the emotion, then they ought to behave in some way. So, for example, some people may believe that feeling an increased heartbeat, higher body temperature, and the urge to fidget constitutes feeling 'anxious'; and they may then associate 'anxiety' with the existence of a reason to leave the situation they find themselves in. There is a double meaning of anxiety here – it refers to both the sensation, and the behavioural disposition; the concept of 'anxiety' links them together. Anxious people sometimes find it helpful to reconceive anxious sensations that do *not* represent a reason to leave the situation as 'excitement' (Brooks, 2014). Chimps, without conceptual representation, do not have this kind of difficulty.

Potential confusions aside, emotional concepts are an essential part of human moral reasons-responsiveness. By understanding what it means to feel 'guilty', 'ashamed', 'angry', 'lonely', etc. humans can more precisely respond to moral reasons. Emotional concepts are the first

[85] Demonstrated by debates about the nature of emotional concepts and whether they are genetic or constructed (Barrett, 2017).

step towards codified moral theories, and in understanding, and subsequently acting upon, moral reasons' content.

Conceptual capacities allow humans to respond to moral reasons about the interests of others more precisely too. Humans not only respond to other minds, but conceptually represent them. Humans have beliefs about the beliefs of others (see 4.5.) and can use these to better respond to moral reasons. Better theory of mind allows humans to interpret and understand when others have interests that we might have some moral duty to promote, and what those interests might be. For example, my conceptual theory of mind allows me to infer that my pasta-loving friend will gain more joy from my pasta-themed gift than my pasta-ambivalent friend. This inference allows me to better respond to moral reasons about (say,) maximising welfare. Humans can do other things with a conceptual theory of mind, too, such as, for example, reconceptualising punishment as rehabilitative or as involving only a limited amount of retributive violence. Rehabilitative punishment practice responds to a more precise moral reason by filtering out self-interested reasons for maximal retribution. Chimps' punishment practice, in comparison, responds to moral reasons more vaguely and less consistently.

Finally, conceptual capacities improve normative guidance by enabling humans to formulate and express moral laws and consistently put them into practice. Normative guidance reaches its fully fledged form in normative concepts. Self-interested motivations can be overcome with reference to rules. This can be done, furthermore, without experiencing moral emotions at that time. The moral reasons-responsiveness is performed through rule-following alone, taking emotions out of the picture. But, if this evolutionary theory of morality is right, their influence is still keenly felt in that picture. (Curry, 2016; Graham et al., 2011; Greene, 2017; Haidt, 2001; Kauppinen, 2013). Humans can do all of this, furthermore, in complex, largely artificial physical and social environments. Understanding the wrongness of turning off the electricity at the hospital to keep it on in a mansion requires a complex and abstract conceptual understanding of both the situation and the moral reasons involved.

Humans have conceptual capacities to respond to moral reasons, but these do not work alone – they work primarily by enhancing the psychological capacities that humans inherited. Conceptual capacities are, combined with humans biological and psychological capacities, adequately moral reasons-responsive. While it may be possible in principle to be adequately moral reasons-responsive without possessing conceptual (or even psychological) capacities,

in the absence of further examples, we ought to put some stock in conceptual capacities as the proven means of developing sufficient moral reasons-responsiveness.

8.6. Plot Twists in the Evolution of Human Morality

I divided the evolved capacities for responding to moral reasons into three. There are *biological capacities*, which provide dispositions to help others (especially close relatives) and *drive* organisms to unreflectively act cooperatively. There are *psychological capacities*, such as moral emotions, theory of mind, and normative guidance, which, enable animals like chimps to respond to a greater range of more precise moral reasons than animals with only biological capacities. Finally, humans *enhance* psychological capacities with *conceptual capacities*. Humans use all three together – one 'source' of human moral behaviour is biological capacities, most obviously in things like parental care. But in the modern day, humans usually respond to moral reasons psychologically, i.e., through defeasible, representative psychological mechanisms like emotions, rather than inflexible biological imperatives. Finally, humans represent moral reasons conceptually, allowing them far greater precision, flexibility, and understanding in responding to them.

We have the story of how human morality evolved. We also know, vaguely, the developmental pathway of morality: biological capacities came first and are simplest, then psychological, and finally conceptual. But there is still one more mystery: what are the conditions required for conceptual capacities to develop?

I have outlined the evolutionary explanations for biological and psychological capacities, which straightforwardly promote genetic and group fitness. The conditions required to develop biological capacities seem relatively straightforward – they arise as adaptations that enhance fitness by solving game theoretic problems in environments (Allchin, 2009; West et al., 2007). The conditions required for psychological capacities to develop are more complex. But there are several evolutionary explanations on offer because many animals have some level of psychological capacities, and variation in their functioning and evolutionary histories offer valuable insights into the conditions under which they evolve. Psychological capacities

plausibly arise as cultural and genetic adaptations for social[86] animals who face complex social problems (Joyce, 2006; Kitcher, 2011; Sober & Wilson, 1998; Tomasello, 2016).

But the conditions for the development of psychological capacities is still far less complex than the conditions for the development of conceptual capacities. Language, at least conceptual language like humans use, has only evolved once. Identifying the conditions under which it evolved is therefore statistically unlikely. There is only one data point, so the statistical effects of potential conditions cannot be isolated. One unknown variable should not immediately lead to despair, but compared to other capacities there is a lack of evidence about the conditions under which conceptual capacities evolved. Despite this, the evolution of conceptual capacities (and language in general) is obviously important and subject to much interest. So, there are several theoretical explanations for why humans alone developed into conceptually competent moral agents.

These theories centre on 'proto-humans' (sometimes called 'hominids') – the distant ancestors[87] of humans who occupy the branch of the evolutionary tree after it split from apes, but before it reached modern homo sapiens. So, under what kind of conditions did proto-humans develop the conceptual capacities that allow them to be uniquely specialised in responding to moral reasons? There are a few candidate explanations.

One explanation is that changes in the climate led to greater rewards for cooperative behaviour and greater punishments for acting alone. "There must have been some environmental difference between early hominins and their great ape relatives, a difference that, perhaps in conjunction with some relatively minor phenotypic difference, initiated a diverging trajectory. One possibility is increasing climatic variation." (Sterelny, 2014, p. 73-74). Higher stakes environments *obliged* hominids to be co-operative. Most think that changes in the climate, or another kind of fortunate mutation or development, kicked off a kind of 'feedback loop' (See Sterelny, 2014 for a detailed account of feedback loops in the evolution of cultural practices in hominids) that led to proto-humans becoming increasingly cooperative and interdependent. Joyce suggests that increasingly large groups were evolutionary pressures to develop language because of the time it would take to maintain

[86] I.e., animals with adaptations to have biological capacities for moral reasons-responsiveness, live in (loose) collectives, and have cognitive capacities like representation.
[87] Though some proto-humans, like Neanderthals, went more or less extinct, they share most of these features.

relationships (Joyce, 2006, p. 90), the so called 'gossip hypothesis' (Dunbar, 1996). Kitcher (2011, p, 68) suggests the larger brains of proto-human babies could have led to more intensive child-rearing and cooperation and that competition between groups of hominids made language and greater cooperation more beneficial (2014, p. 106-8). Sterelny also suggests that group competition and collective defence could have been instrumental in kicking of the feedback loop that results in language (Sterelny, 2014, chapter 6).

The evolutionary conditions for the development of conceptual capacities could have been many things, and there is little certainty about it. The general conditions were that proto-humans were *ready*, in a chimp-like state, to learn; and that then they gradually pushed towards greater and greater moral refinement, to the grand conclusion of conceptual competency.

8.7. Evolving Artificial Moral Agents

So, where does this chapter leave us in terms of designing evolutionary artificial agents? I reiterate that human capacities for moral reasons-responsiveness are *sufficient* for adequate moral reasons-responsiveness, but I have not argued that they are *necessary*. Likewise, the various capacities of moral reasons-responsiveness in nature are sufficient for various levels of moral reasons-responsiveness, but they are not necessary for moral reasons-responsiveness. Perhaps psychological and biological capacities can develop to respond to a greater range of more precise moral reasons. I see no reason to think that it is impossible for an agent with only biological or only psychological capacities to be adequately morally reasons-responsive. However, in the evolutionary history that we know about this has not happened.

While it may be possible for an agent to be adequately moral reasons-responsive without conceptual capacities, a designer who wants to be most confident about designing adequately moral reasons-responsive agents ought to aim to design agents with all three of the capacities discussed here. The possibility of this will be discussed in the next chapter, but I note here that from in evolutionary history the development of biological and psychological capacities is common – there is therefore good reason to think that an evolutionary artificial agent could develop these capacities.

The obvious means of doing so is for the artificial agent to develop first biological, then psychological capacities. Biological capacities are, given their ubiquity in nature, relatively easy to simulate. Psychological capacities pose more difficulty, especially the chimp-equivalent psychological capacities I focused on. But, given the mutation of biological capacities and their role in incentivising and facilitating greater social integration from a fitness point of view, it is reasonable to expect that, unless something goes badly wrong, artificial agents can evolve psychological capacities. Conceptual capacities pose the greatest challenge, although recreating some of the conditions described in 8.6. may aid a designer who can do so without undermining autonomy.

Finally, any agent that develops adequate moral reasons-responsiveness through evolution is also likely to satisfy the nonhistorical conditions of responsibility. These are standardly 'epistemic' and nonhistorical 'control' conditions. The psychological skills of representation and normative guidance imply the ability to meet these conditions. Possessing normative guidance is, as far as I can tell, a key component of meeting the nonhistorical control condition of responsibility. Since normative guidance seems to track the general ability to be precisely reasons-responsive or align one's higher-order and lower-order desires. The epistemic condition can likewise be taken to be met by any evolved agent with adequate moral reasons-responsiveness – since the psychological capacities involved already require the ability to anticipate consequences and predict reasons. I will not discuss these nonhistorical conditions of responsibility further – though I think the relationship between normative guidance and nonhistorical conditions for responsibility is worthy of further research. I think it is reasonable to assume that an agent which evolved psychological capacities to respond to moral reasons meets the nonhistorical conditions for responsibility. One consequence of this is that, should it be possible, an agent that evolved the ability to adequately respond to moral reasons using biological capacities alone (the 'ultra-bee') would be unable to be responsible because it would not be able to represent its desires. This seems intuitive to me, but in any case the developmental pathway I am suggesting does involve psychological capacities, so this point should not be of further concern.

9. Evolving Artificial Moral Agents

9.1. Introduction

In this thesis so far, I have argued that artificial systems are generally not moral agents because they cannot simultaneously satisfy the conditions for autonomy and be adequately moral reasons-responsive. Humans have autonomy-promoting histories largely because they evolved. I have argued that the reason artificial systems normally do not is because their actions have agential causal sources, and they do not have sufficient non-agential causes, like evolution, to be autonomous.

In this chapter, I will claim that artificial life systems can be the exception. I argue for this by outlining how artificial life systems can have autonomous moral capacities by developing through open-ended evolution in a complex simulated environment. Therefore, artificial life systems can succeed where other artificial systems fail and have the best chance out of any type of artificial system for becoming a moral agent.

Artificial systems like the Moral Decision Machine in chapter 6 can be adequately moral reasons-responsive. But they cannot achieve this without an autonomy undermining history. Artificial life systems can have autonomy-promoting histories, but if so, they cannot develop adequate moral reasons-responsiveness through moral deference. If they are to develop adequate moral reasons-responsiveness, they must follow a different path. This path was outlined in the previous chapter – human moral reasons-responsiveness evolved. Since artificial life organisms are evolutionary, they may be able to follow the same path. I argue that artificial life organisms have a high chance of evolving a decent level of moral reasons-responsiveness equivalent to chimp moral reasons-responsiveness, and a lower, but still significant, chance of evolving to be adequately moral reasons-responsive.

The structure of the chapter follows. In 9.2. I explain what artificial life is, how it is normally used, and what it can do. In 9.3 I describe how artificial life organisms can have meet the historical conditions for autonomy. In 9.4, I discuss how autonomous artificial life organisms evolve to respond to moral reasons through biological and psychological capacities. In 9.5, I offer a proposal for how autonomous artificial life organisms' moral reasons-responsiveness may be useful outside artificial life simulations. In 9.6, I discuss the most challenging

capacity necessary for adequate moral reasons-responsiveness: conceptual capacities. I offer some rough sketch of how artificial systems may develop conceptual capacities.

9.2. Artificial Life

In this section, I give an overview of what artificial life is and how it is normally used. An understanding of how artificial life systems work and what they can do is necessary for the following sections, which describe how artificial life systems can be both autonomous and morally reasons-responsive, to proceed smoothly.

'Artificial life' is the design and study of artificial systems that use evolutionary principles and allow a program to evolve (Kim & Cho, 2006). They are often complex simulations of generations of pseudo-genetic algorithms or 'organisms'. Artificial life is to be distinguished from 'genetic algorithms' in general (Suggested as a means of developing artificial moral agents by Muntean & Howard, 2016). Genetic algorithms can be designed to solve specific problems and use evolutionary forces to do it. Artificial life systems, in contrast, are used to reveal facts about evolutionary forces, to develop artificial organisms, or to test hypotheses about evolutionary history (Taylor & Jefferson, 1993). Artificial life systems can yield results that can inform mathematics, evolutionary theory, population dynamics, sometimes healthcare (as virus and bacteria can be simulated decently well), and xenobiology. Artificial life systems are not typically used for commercial applications like genetic algorithms in general are. They are more like aquariums (or terrariums) where researchers grow artificial agents, like little fish, to learn about the world. Artificial life systems are not themselves evolutionary algorithms or neural networks. But each artificial life organism may be a neural network or evolutionary algorithm.

The Artificial life organisms I will discuss operate within an isolated digital environment – their 'fish-tank', to keep the analogy rolling. The combination of organisms and environment – the aquarium seen as a whole, both tank and fish, is the artificial life system. Artificial life environments can have varying similarity to (both past and present) Earth. Artificial life designers need not aim for environments that are historically realistic. Even if they did have that aim, increasing historical realism typically means increasing complexity, which isn't always practical, necessary, or affordable. Artificial life designers' goals can often be satisfied by a simpler, less realistic system. Sometimes their goals are opposed to realism,

such as the goal to test non-Earth like systems – for example, developing an artificial organism that would thrive on Mars or is based on silicon instead of carbon.

Artificial systems can be embodied the traditional sense, or 'hard' or they can be digital, or 'soft' (Aguilar et al., 2014; Kim & Cho, 2006; Sullins, 2006). A robot is a 'hard' embodied artificial system: it has a body, appendages, and can interact with the same environment that humans interact with. A robot's actions and interactions are governed by its program, which is 'embodied' *by* the robot, but may not be physically within the robot itself, and instead could be in a server (or other device) elsewhere. Artificial life systems *can* be robotic (Eiben, 2014; Steels & Brooks, 2018). But they face many challenges in reproducing and adapting to the environment. Further 'hard' artificial systems are 'wet' – made of synthetic biological material (Hanczyc, 2020).

A 'soft' artificial life organism is not similarly embodied, it operates within a digital environment and cannot interact with the human environment. A 'soft' artificial life organism that "would thrive on Mars", for example, would be one that thrived in a digital simulation of the environment of Mars (with the corresponding gravity, atmosphere, land features, etc.). Despite not being embodied in the traditional sense, 'soft' artificial life organisms are still able to be agents, as they operate as an agent within their digital environment. They are also embodied in a weaker sense, in that their program controls a digital body that interacts with a digital environment.

I will focus on 'soft' artificial life for the following reasons. First, robotics systems are not a viable route to autonomy – they are constrained in that they must be specifically designed to interact within the physical environment, and they cannot mutate truly open-endedly. Both of which suggest that, while useful on their own terms, they are dead-end for this chapter's aims. 'Wet' artificial life organisms may ultimately develop into artificial moral agents. However, it is in an early stage and there may (and probably will) be unforeseen difficulties. Synthetic biology is currently at the stage of developing individual cells, rather than complex agents. Digital artificial life, in comparison, already generates agential organisms. Though I will not justify the decision to focus on digital artificial life further: excluding wet artificial life here is in the interest of maintaining a reasonable scope for the chapter.

Here, I focus on 'soft' digital software agents. For now, I will avoid the challenge of embodiment for artificial life systems. If you can make a digital artificial moral agent, then it ought to be possible, although more difficult on a practical level, to make an embodied

artificial moral agent using the same techniques. Though I do not discuss methods of embodying digital moral agents, I do discuss, in 9.5 I discuss some ways of putting 'soft' artificial life morality to use.

With the scope set, let us zoom in on 'soft', digital artificial life systems. These systems are a simulation of both organisms and environment. In the field of artificial life, there are several platforms for simulating artificial life systems using ordinary computer hardware such as Tierra, (Ray, 1993) and Avida (Ofria & Wilke, 2004). These, as mentioned, are normally used for research in various domains. The simulated environment produced with these platforms can be simple or complex, but most of the time it is rather minimal, consisting of basic environmental features, such as 'food' 'danger', 'competitors', 'mates' and 'co-operators'. The environment is populated with agents – artificial life organisms -- that react to those features. But before exploring these complex systems, it is worth describing a simple one to illuminate the basic principles involved.

One of the simplest types of artificial life system is Conway's 'Game of Life'[88] (Berlekamp et al., 2003). It is simple enough that it can be simulated by hand. The Game of Life's environment is a grid, there are only two values: on and off. Off is represented by white, and on by black. The rules for time-progression are the following:

1. Any on cell with two or three on neighbours stays on.
2. Any off cell with three live neighbours turns on.
3. All other on cells turn off (those with more than three or less than two neighbours).

Conway's Game of Life's starting configuration is with some cells on and some off. With each 'tick' forward, the game's rules are implemented, and the outcome is then subject to the rules again on the next 'tick' forward. See the opposite page for an example. Game of life systems tend to go on for a period of ticks before settling into stable patterns or dying out completely. In the image opposite, 12 is a stable pattern – all cells have two neighbours, and therefore each tick that goes by they all remain on, and no new cells turn on. For easier reading, I refer to the 'ticks' of state transitions as 'time'.

[88] Though Neumanns' cellular automata are accepted as the very first artificial life system.

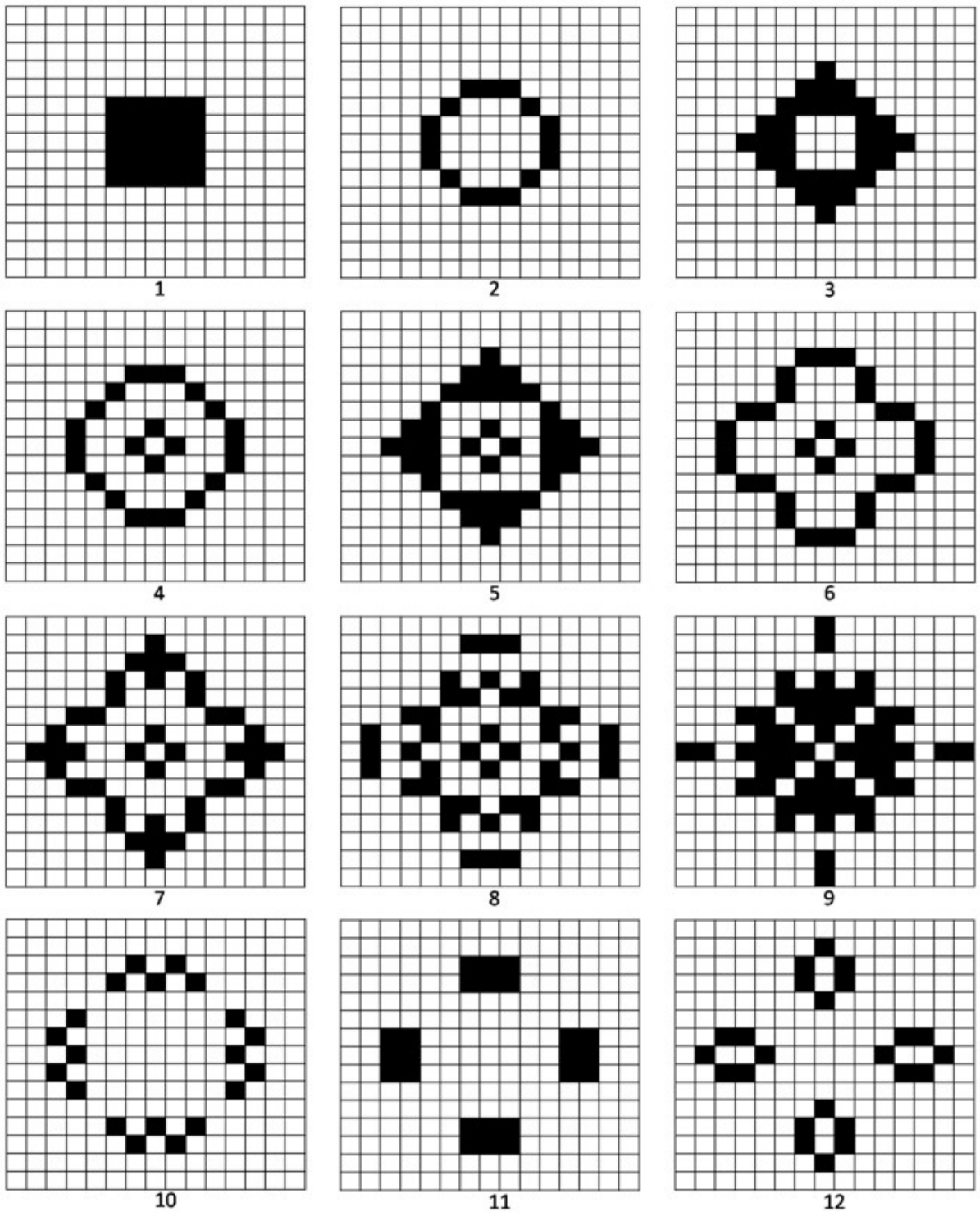

Figure 1 The evolution of Conway's Game of Life for a 5x5 'alive' cell initial configuration. The black cells are 'alive' states and the white cells are 'dead' states. The first 12 evolutions are shown, and the last one is a steady state.[89]

[89] Figure is taken from Clarridge (2009)

The Game of Life, simple as it is, demonstrates interaction between agent and environment. The environment is both the lattice that the cells operate in and the rules of the game. Humans have a similar relationship with their own environment: natural laws (such as physical, logical, and mathematical laws) play the role of the 'rules of the game', and the physical world plays the role of the lattice with on and off states. In the more complex artificial life systems about to be discussed, artificial life organisms relate to their environment in the same way. For complex artificial life systems, just like in the game of life, designers can make decisions about the variables involved, the substrate on which variables occur, and 'rules' (or, more evocatively, 'natural laws').

Complex artificial life systems use specific environments to generate different types of (digital) organisms[90], with different behavioural dispositions, different types of beliefs and motivation, and who sense variously complex or simple values in their environment. The rules of artificial life systems are, of course, much more complex than the Game of Life. But certain types of rules will always be necessary, at least if you want to create an artificial life system that is at least somewhat 'life-like' and develops complexity through natural selection (which is the aim here, at least). There needs to be an 'evolutionary principle', the conditions under which organisms can reproduce (either sexually or asexually). There should be rules that generate something like mutation, for example, rules that cause random modifications to organisms as part of the reproduction process. These mutation-like rules serve a similar kind of 'function' as mutation in biological evolution - they allow new fitness promoting strategies to be developed. Finally, there also need to be general physical rules that allow organisms to exist and act.

Artificial organisms develop behavioural strategies that reflect their environment. An artificial life system with an environment similar to our ancestors' environment, where there is a clear reproductive benefit to co-operative tendencies, produces artificial organisms that co-operate with one another. But an artificial life system where the rules of the environment lead to distinctly zero-sum situations, offering no reproductive benefit for co-operation, produces artificial organisms that do not co-operate. Often artificial life systems tend towards

[90] These organisms, as will become clear, unproblematically satisfy the conditions of agency described in 3.2.

an evolutionary equilibrium – where either one strategy is dominant or where there is a balance of strategies.

Here it is useful to bring in an example. The Digital Evolution Lab at Michigan State University use an artificial life simulation platform called 'Avida'. Avida is an artificial life simulation that focuses on the 'organism' level, as opposed others that focus on the 'species' level or even 'ecosystem' level (Such as RevoSim, see Garwood et al., 2019). Avida, like Conway's game of life, takes place on a lattice. Some cells within that lattice are artificial organisms. In Avida, organisms contain a lot more information than simply 'on' or 'off'; they consist of a 'genome' – a series of instructions that is executed over time. These genomes can mutate randomly when they reproduce by randomly adding (or deleting) a new instruction from a list. These mutations lead to interaction between diverse organisms and thus facilitate the development of various behavioural strategies.

The typical Avida (Ofria & Wilke, 2004) artificial organism lifecycle is like this: first, the organism is 'born' – it arises from two existing organisms that sexually reproduce (when they have the disposition and energy to do so – a situation determined by their genome and their interaction with the environment). The baby organism then begins executing its instructions – sometimes competing, sometimes cooperating, sometimes feeding, sometimes reproducing, sometimes moving, etc. It executes its instructions linearly according to its genome, which it inherited as a combination (with possible mutations) of its parents' genomes. As it acts, other organisms interact and respond to its actions. If it successfully reproduces, it passes on its genome. Finally, upon reaching the end of its lifespan (which is determined arbitrarily) or depleting its energy, it dies. Tragically, when Avida organisms die they vanish from the environment completely – fortunately the user benefits from a bird's eye view, and may pause time, or jump forward or backwards within it, to check up on the organisms' progress.

Avida organisms can develop complex capacities and overcome various obstacles in that development. Consider the evolution of artificial systems to solve computational problems in Avida (Ofria & Wilke, 2005). The artificial organisms developed the capacity to solve a set of problems that required the use of the logical function EQU, using a set of simpler capacities. The simpler capacities were not themselves logical functions but needed to be placed together in the right way. The artificial organisms did not *need* to solve the more complex EQU function – they had alternative problems to solve, although those problems were ultimately less efficient. By competing, the artificial organisms evolved to solve the

most complex function. The authors aimed to show that evolution could lead to complex problem-solving that itself involved maladaptive more basic capacities. To solve EQU, the artificial organisms needed to evolve a series of simpler capacities that, until the last piece of the puzzle was placed, made the organisms less competitive.

Artificial organisms can evolve capacities. Sometimes these capacities can be sophisticated. Sometimes complex capacities arise seemingly improbably, with a strong resemblance to the evolution of biological organisms' capacities. So far, artificial life systems like these seem like a promising candidate for the simulating the evolution of moral reasons-responsiveness. However, evolving to solve computational problems may seem to be some distance away from solving distinctly *moral* problems or responding to *moral* reasons.

Clune et al. (2011) shed some light on Avida's progress toward *moral* reasons-responsiveness. They describe how a version of Avida was used to develop artificial organisms with kin-based 'altruistic' dispositions. The ability to help (at a cost) those with similar genomes to oneself was added to the list of possible mutations, and the simulation was run to see if the mutation increased reproductive fitness. Compared with the original Avida organisms, and those who helped every other organism indiscriminately, the kin-identifiers gained a fitness advantage. This reproduces the effectiveness of kin selection and demonstrates that artificial life organisms can be subject to evolutionary dynamics such as kin selection.

Since then, co-operative strategies have been robustly successful in artificial life systems (Aguilar et al., 2014). This proves that co-operation is selected for by evolutionary dynamics, but it also shows that artificial life systems can contain organisms that can develop traits through 'natural' selection. The disposition to identify and favour kin also closely resembles biological capacities for moral reasons-responsiveness, so these artificial life systems must bear some interest for the artificial moral agency researcher. However, we should not go as far to say that the kin identifying organisms in Avida are on par with even the simplest animal morality.

One reason for this is autonomy-undermining human intervention. Animals on Earth evolved in an uncontrolled, natural environment. The evolutionary dynamics that governed their development are those that naturally emerge from physics. The same goes for their environment, their competitors, and the complexity of their environment. Avida's artificial organisms did not share this liberty. Their environment is determined by the whims of the

designing team (who attempt to control the environment to test a single variable – such as the value of kin selection in evolutionary processes). The artificial organisms follow a set template with set mutation possibilities. Their make-up, possible behaviours and environment are all constrained by the designers of the system. Despite the distance a designer has from their artificial organisms, their causal fingerprints on the organism are still clearly visible. Consequently, even if these artificial organisms have moral capacities, they do not develop them autonomously (a concept used in the same way here as in chapter 6[91]), that is, they only develop them because of the causally invariant intervention of the designers. To design autonomous artificial organisms, the designer's fingerprints need to be scrubbed from the artificial organisms' causal history. Doing so requires an artificial life system that lacks causally invariant intervention from designers. In the following section (3), I will argue that this is possible, and that artificial organisms can be autonomous when they develop in some types of open-ended artificial life systems.

The second reason that the Avida organisms just described are only a limited success from the artificial moral agency perspective is that they do not have *adequate* moral reasons responsiveness. Their moral capacities do not involve the psychological capacities for moral reasons-responsiveness, let alone concept capacities.

What artificial life has achieved so far is substantial. A route to the development of complex capacities has been sketched out and implemented in a limited way, and this route may, given the right implementation, lead to fine-grained reasons-responsiveness. So, the first condition of moral agency, in terms of reasons-responsiveness, looks solvable for artificial life organisms. However, other artificial systems, such as the Moral Decision Machine in chapter 6, are also sufficiently morally reasons-responsive. The central obstacle for artificial moral agency is the combination of autonomy and moral reasons-responsiveness. In the following section, I explain what would be needed to create autonomous artificial organisms. That, unfortunately, is the simplest step. The next step is to create autonomous artificial organisms that can develop adequate moral reasons-responsiveness. Such a goal is distant, but, I believe, achievable.

---

[91]Some in artificial life and elsewhere distinguish between 'behavioural' and 'constitutive' (relating to autopoiesis) autonomy – the focus here is reasonably on 'behavioural' autonomy. But I do not use this distinction, as it seems to me that 'constitutive' autonomy may be relevant to an organism's meeting historical conditions of autonomy.

9.3. Autonomous Artificial Organisms

Consider an artificial life organism compared to a machine learning algorithm. How close are they to satisfying the historical condition of autonomy (from chapter 7)? Many machine learning systems achieve their functions in a way that appears independent to the designer or user. The influence that the designer has on the machine learning system is in defining its function: programming the ways in which the machine learning program can learn, controlling its measures of success, its paradigmatic examples, etc. As we know, these systems are not autonomous because they have been designed in an autonomy undermining way – i.e., their actions have their causal sources in other agents' actions.

The artificial life organisms that considered so far have generally equivalent histories. The artificial life organisms solve functions using evolutionary dynamics, but their goals, general solution-building toolkit, and operation is still determined by the designers. These types of artificial life system do not hold any advantages, in terms of autonomy, over machine learning systems.

However, I will outline how to design artificial life systems that do have the appropriate history. The artificial life systems considered so far have had their goals defined by their design – in the Avida examples from the previous sections, those goals were to solve EQU and to test whether kin-identification promoted fitness. In the first case, the goal determination is clear. The designers use the environment of the artificial life system to place a series of environmental challenges – after these challenges are met, the artificial life system may evolve to be more efficient, but it cannot evolve to perform different tasks. In the second case, the goal-determination is more subtle. But, nonetheless, the artificial life system will reach an evolutionary equilibrium that provides the result the programmers are looking for. The designers do not add any additional complexity to the environment that may prevent this equilibrium – they set up the environment such that the artificial life organisms cannot evolve beyond the state that offers an answer to the hypothesis. This can be contrasted with natural evolution, where the environment is constantly changing and in which evolutionary equilibriums have tended to be short lived. There are some exceptions, of course, some fish have not made significant evolutionary advances for millennia – but humans were biologically evolving, and continue to evolve, in the present day. The question is, then, can artificial life be subject to highly complex, open-ended evolution that resembles natural evolution, rather than the closed, predefined goals of current artificial life systems?

Offering a positive answer to this is a central goal of artificial life research, Mark Bedau notes that "the aim of many Alife models is an open-ended evolutionary dynamic that is forever far from equilibrium." (Bedau, 2004). An artificial life system can be run with a highly complex environment and an open-ended evolutionary process. Let us take these in turn. First, the environment must prompt enough pressure for almost all evolutionary equilibriums to be unstable. There are design-focused ways of doing this, such as by initiating extinction events every time an equilibrium is reached. But this evolutionary pressure will also occur if the environment is complex enough. One way of adding to the complexity of the environmental system is to have biomes – or modifications on the natural environment – such as a weather-like temperature system, a tectonic-like land movement system. Such complexity-inducing modifications need not be identical to Earth's, or resemble Earth in any fine-grained sense, but they ought to be recurring, rule-governed processes that change which organisms are most fit. For example, if there is a temperature gradient throughout the artificial environment, artificial organisms will evolve to solve problems in high and low temperatures differently. The artificial organisms, like natural organisms, can be expected to adapt to their environmental niche and develop into competing artificial species – one that is adapted to cold environments, and another that is adapted to warm environments. This diversity will ensure competition – the 'heat-adapted' artificial organisms will compete with the 'cold-adapted' ones. Another way to achieve a similar type of complexity is to have 'sea' and 'land', again requiring different types of adaptation and ensuring competition (and destabilising evolutionary equilibriums).

Another way of conceiving of environmental complexity is through the variety of available evolutionary niches. For example, an artificial environment that can support both carnivorous and herbivorous behaviour. Or that can support both 'agile' and 'strong' artificial organisms. Again, the goal is to destabilise evolutionary equilibriums by maintaining a multi-pronged, interactive competition, just like the competition of our own evolutionary history. Call this the 'environmental complexity' variable.

High environmental complexity would mean little if the organisms involved can only evolve into a few preset templates or can only have a few types of solutions to problem. In natural evolution, organisms mutate on a fine-grained level which can result in complex adaptations – after all, our shared evolutionary ancestor is a simple group of cells, which eventually evolved to be organisms as complex as us. This involved emergent capacities that initially seemed far out of reach. An artificial organism, if constrained to a set of predefined

instructions (such as, for example, the logical operators), will not develop an 'eye'. However, an open-ended evolutionary system that does allow for fine-grained adaptations from which complex capacities may emerge from can be created. To do so, one needs a working system of physics in the artificial environment. With an evolutionary process that depends on a theory of physics, artificial organisms can evolve open-endedly and develop new morphological capacities, including, at least in theory, brain-like capacities. The theory of physics enacted within the artificial system does not need to be as complex as our own physics (that would be impossible), but at least fine-grained enough to be equivalent to the cellular level. With a working cellular physics, artificial organisms can evolve freely by adapting cells – they are still constrained by the physical system (as natural organisms are), but they are no longer constrained by the ingenuity of the programmer. Call this variable 'physical complexity'.

If an artificial life system has both physical and environmental complexity and there are no further arbitrary constraints designed into the environment, then it will evolve in a way that resembles natural evolutionary history. This, should it occur, would mean that the artificial organisms' actions would have their causal source in evolutionary forces. The designers' actions, despite conferring a function, would have a weaker causal relationship with the organisms' actions than their evolutionary history. Such a system evades the autonomy-undermining design: the artificial organisms have a similar (autonomy promoting) history to natural organisms.

It should be further noted that the competition between artificial organisms, although they are all agents, should not be taken to problematically affect their history. In the same way that natural organisms competed and, despite being agents, did not undermine one another's causal history. All natural organisms have a high degree of interaction with other agents, but that interaction does not normally constitute autonomy undermining involvement – artificial organisms are in the same boat.

A final point is that the designers still make certain interventions in the artificial life system – they will (unless they can create truly complex natural-like physics) design the simplest cellular or subcellular organisms. Likewise, they will have to design the environment, and this may lead them to have some effect on *how* the artificial organisms will evolve. There are two options here. First, bite the bullet and hope that these design elements will not be sufficient to undermine the autonomy of the artificial organisms. Second, automate the

processes of environment and agent generation – indeed, such automation will probably be necessary, simply because designing viable artificial organisms, even simple ones, is a computationally involved process that would be challenging to do without some automation tools. An iterative process can 'design' simple organisms through an independent evolutionary algorithm – this algorithm will not itself be autonomous, it will be designed with the aim of producing subcellular or cellular life that can support sufficient physical complexity, or with producing an environment with sufficient environmental complexity. This process can ensure that the designer's interventions have sufficient causal variance to support the autonomy of the artificial life organisms. If the agents and environments are designed in this procedural (i.e., with the designer defining the general procedure but not the output) way, then the artificial life system will be unlikely to have the designer's actions as their causal sources.

There are many difficulties with achieving environmental and physical complexity. There would be many practical challenges, including computing power. A philosophical difficulty is that the procedural generation of an artificial life system may allow for one last gasp for autonomy-undermining intervention: choosing the simulated world. Suppose that you generate 10,000 environments and 10,000 agents – if you intervene by selecting the agents and environments to be simulated according to certain criteria (say, chances of the artificial organisms developing brains), then you may be undermining the autonomy of the artificial life system. The autonomy supporting approach is to choose them randomly.

Achieving environmental and physical complexity in this way should lead to the evolution of autonomous artificial life organisms. They would far outstrip, in terms of autonomy, a machine learning system that solves a particular function. Essentially, the way to think about this problem is this: as artificial life designers are like Gods. They can choose to be interventionist Gods (like in the examples of artificial life in the previous section), or they can choose to be non-interventionist Gods. Intervening infringes on the autonomy of the artificial organisms, so we ought to take pains to be non-interventionist, using the physical constraints of possible systems of physics, procedural generation that mimics natural processes through randomness, and by refusing to make decisions about the nature, lives, and goals of the organisms.

These artificial organisms would be autonomous, and can, at least in principle, evolve complex capacities. There are yet two further problems before this can be considered a

serious methodology for developing artificial moral agents. Firstly, how do we develop artificial organisms that have moral reasons-responsiveness without impinging their autonomy, and secondly, even if we do develop these artificial organisms, can their moral capacities ever be applied to the natural world?

I believe that the first problem is less problematic than it appears and is solvable with our current technology. However, the second problem is relatively problematic. Although artificial moral agents that are constrained to their artificial environment ought still, in my mind, be considered a 'win' for artificial moral agency. I will discuss both problems further in the following sections.

9.4. Artificial Organisms with Moral Reasons-Responsiveness

Suppose that we have an artificial life system with sufficient physical and environmental complexity for artificial organisms to satisfy the historical condition of autonomy. The capacities they develop will inevitably resemble natural organisms' capacities – there is highly likely, for example, to be a perceptual system, an internal monitoring system, a system for movement, a system for making decisions, and so on. Would they also develop *moral* reasons-responsiveness*?*

In the previous chapter I identified a series of landmarks in the evolution of morality. Humans have met all of these landmarks and most animals have met some too. The first of these landmarks are 'biological capacities' – that is, genetic, or otherwise driven by adaptive mechanisms, dispositions to behave in prosocial ways. They are done unreflectively, involuntarily and instinctively. Psychological capacities, in contrast, require some level of representative ability, and involve a cognitive layer of moral reasons-responsiveness which can facilitate greater moral reasons-responsiveness. Human moral reasons-responsiveness involves both psychological and *conceptual* capacities. Conceptual capacities being dispositions to respond to moral reasons using conceptual representation of goods, aims, and ends. I leave conceptual capacities aside for now. Let us consider what would be needed for artificial organisms to evolve biological and psychological capacities.

We would be faced with an analogous story to the development of biological capacities on Earth. Evolutionary dynamics offer an adaptive benefit to cooperative strategies such as kin selection, reciprocity and mutualisms. An open-ended artificial life environment will

therefore, given sufficient complexity, produce environmental challenges that are best solved by the development of moral reasons-responsiveness. Artificial organisms that mutualistically protect and defend one another will outcompete artificial organisms that do not; just like in natural history. Similarly for groups that can develop reciprocal relationships and advanced cooperative behaviours. Since these strategies are adaptive, we can expect that, once mutated in an open-ended evolutionary simulation, they will (eventually) proliferate.

There are interesting questions about what proportion of complex environments will cause moral reasons-responsiveness to be adaptive. Consider this charge of 'anthropocentrism': humans and other species on Earth evolved to be cooperative because of the unique environmental circumstances of Earth. There are many possible complex environments that do *not* offer any advantage in reproductive fitness from cooperation. Perhaps by being inspired by the history of organisms on Earth, I am ignoring the possibility of non-Earth like environments.

However, I think it reasonable to assume that cooperative solutions will *always* be adaptive to some extent. The payoffs of cooperating seem to apply generally to a wide range of possible evolutionary scenarios. This is demonstrated not least by the ubiquity and variety of cooperative strategies in Earth's history. Insects cooperate, just as fish, mammals, reptiles, even some types of bacteria. So long as we accept the standard evolutionary explanation of moral reasons-responsiveness, we ought to believe that it will arise in most environments that support complex organisms. A secondary reason for believing that moral reasons-responsiveness will always be adaptive is that they reflect game-theoretic mathematical theory. Game theory *proves* that cooperative behaviour is adaptive in all worlds that share evolutionary forces. Of course, in some distant, alien worlds, game theory may be shown to be false, or perhaps innovative, non-co-operative solutions to game theoretic problems can be shown to be realistic. But the ubiquity of cooperation on Earth, and its verification by game theory ought to be sufficient to convince us that all complex evolutionary systems will converge on moral reasons-responsive solutions.

If so, then an artificial life system with complex evolving organisms will develop organisms with moral reasons-responsiveness. They would involve cooperative dispositions, solutions to social problems that require cooperation, and fitness sacrificing between individuals. Note, however, that this is clearly not *human-level* or *adequate* moral reasons-responsiveness. They

are moral reasons-responsive in the same way that a bee is, that is, they have *biological capacities* for moral reasons-responsiveness.

That psychological capacities build on, in some way, biological capacities is the majority view in the literature on the evolution of morality. The most extreme version of this view is that psychological capacities *just are* advanced biological capacities. According to this view, artificial organisms with biological capacities ought to evolve psychological moral capacities (given sufficient environmental and physical complexity).

One way to point to this connection is to suggest that, as those like Rottschaefer do, psychological moral capacities are simply the application of complex cognitive capacities such as representation to biological capacities. There is a well-travelled road, they might argue, between biological and psychological capacities. Therefore, we can trust that an open-ended artificial life evolutionary system that develops artificial organisms with biological prosocial dispositions will also, eventually, evolve artificial organisms with psychological moral capacities.

Essentially, the answer to the question of *how* to evolve artificial life systems that have prosocial representations is that, if our evolutionary theory is right, we won't need to truly understand *how* it happens. If we set-up a realistic enough evolutionary simulation, then it *will* happen. Both phases of biological capacities and more complex moral sentiments, intuitions or other representations should occur over time. They may perhaps occur in unexpected or unanticipated ways but would still facilitate greater responsiveness to moral reasons. Call this strategy the 'biological strategy'.

A milestone development in the evolution of psychological capacities is mental representation. Psychological capacities such as those chimps have involve a source of moral evidence such as emotion or intuition, the capacity for normative guidance, and being embedded in a society of obligate co-operators. These capacities require advanced cognition and mental representation. Emotions and intuitions *are* acts of mental representation. For an agent to have a theory of mind, it must represent the goals and intentions of others. For it to have normative guidance, it must represent the 'command' of its emotion or intuition.

Capacities for mental representation are a plausible bottleneck in the pathway from biological to psychological capacities. You might think that the development of these advanced cognitive skills requires perhaps a lot of luck, or perhaps a very specific, and far from guaranteed, set of environmental conditions. One suggestion, therefore, would be to design

artificial life organisms *such that* they initially possess these representative abilities. Artificial life organisms could get a boost, by innately possessing representative abilities because they could develop psychological capacities immediately. *Some* biological prosocial dispositions may still evolve, in cases where context sensitivity would be a detriment to the overall fitness of the organism (such as, say, for the disposition to care for babies), but the bread and butter of the evolution of sociality: game-theory based solutions to co-operation and competition problems, would surely take advantage of the representative capacities on offer. Call this strategy the 'psychological strategy'.

There are two drawbacks to the psychological strategy in comparison with the biological strategy. First, designing capacities for representation can be difficult, may have unintended consequences (for example, representative capacities would probably need to have a 'scope', meaning that an artificial life organism had the innate disposition to represent certain *types* of objects, and there is no guarantee that this scope will be adequately defined by a designer). Second, by taking a heavier hand in the design of the artificial organism, a designer may undermine the organisms' autonomy.

In this, the biological strategy is less certain to achieve psychological capacities but more likely to produce autonomous organisms; while the psychological strategy is more certain to achieve psychological capacities but more likely to undermine autonomy. I am reluctant to anticipate the success of these either strategy. Both depend on successfully developing a complex digital physics and having a large amount of computational power. But it seems to me that either strategy could be effective in producing autonomous moral reasons-responsive (to the level of chimps, or so) artificial agents.

9.5. The Moral Module

Instead of speculating further, I will turn to two anticipated challenges that lie between these artificial organisms and artificial moral agency. First is the challenge of transferring artificial organisms' to the 'real' world, and second is developing *adequate* moral reasons-responsiveness.

Imagine artificial organisms that have been produced by an open-ended complex evolutionary simulation. They cooperate and perhaps they are obligate cooperators. They have a kind of representative capacity that allows them to adapt in their lifetime to novel

situations. They can respond to theft of a new resource as if it is theft of a resource they already know – they can punish repeat offenders or offer displays of shame and humility when they have acted in their own interest at the cost of the group. They do so via beliefs or emotions which serve as motivators for, for example, norm enforcement, displays of contrition, and group-aim promoting actions. In short, they are like artificial chimps.

These artificial chimps, although they and 'real' chimps are equally autonomous and morally reasons-responsive (which would be a big step forward), only respond to their artificial environment. So, they seem to have few uses in the 'real' world. If making artificial moral agents requires cutting them off from the human world, then the project of developing them is restricted in impact (and disincentivised).

However, there may be some ways of putting artificial chimps to work. Their capacities are likely to be accessible from a designer's God-like perspective. A designer can gather lots of data about which organisms have evolved in which exact ways, how exactly the organisms behave. Crucially, a designer can isolate an individual organism's 'program' and identify the functions of different dispositions. In a system like Avida, this is simple – as organism's programs are combinations of instructions drawn from a finite set, so any organism's program can be isolated and functionally decomposed. If the artificial life system is environmentally and physically complex, it will not be so simple to do this. A designer would need to interpret which parts of the organism performed which function. This interpretation ought to be fairly accurate, however, because the designer would have a great volume of reliable data, including the evolutionary history of each part of the organism.

A designer can interpret from this data the physical realisers and the intentional content of the organism's dispositions, representations, perception, and memory. As the artificial organism would be based on a simpler version of physics compared to our own, the dispositions and representations of the artificial organisms would be simpler by a similar measure, and we would have an easier time identifying which parts of the artificial organism are responsible for moral reasons-responsiveness compared to similar exercises with humans (which have so far been inconclusive). This can again be benefited by a wealth of data and experimental power - allowing moral psychology experiments to be run on artificial organisms at speed.

Having identified the parts of the organism responsible for moral reasons-responsiveness, a designer could then work on translating that capacity to one that can apply to real world situations. The designer would need to reverse engineer the artificial organism's capacities

for moral reasons-responsiveness. If this can be done, then the designer will have something like a 'moral module' that responds to moral reasons. This module can be fed different scenarios (devised within its environment, rather than the 'real' environment) and respond to the moral reasons in them.

It is possible that the artificial chimps' capacity for moral reasons-responsiveness could not be isolated from other capacities. This would complicate things. The goal of reverse engineering the moral reasons-responsiveness would remain, except there would be a lot more capacities that would need to be reverse engineered. A designer would hope for a complete, localised, moral module in a simple artificial organism with adequate moral reasons-responsiveness. But the nature of the artificial chimps is, at this point, difficult to predict without further experimental evidence.

The differences in difficulty between reverse engineering holistic and modular artificial minds reflects a more general tension between complexity and functional decomposability. The more complex the artificial environment is, the more likely it is that the artificial organism will be able to evolve open-endedly and thus develop chimp-level moral reasons-responsiveness. However, greater complexity also correlates with the difficulty of reverse engineering the organisms' capacities.

Whether the artificial chimps have holistic or modular minds, a designer ought to be able to generate a moral module that can respond to the moral reasons in any given situation from the artificial environment. However, the moral module would be unable to parse the real world.

The next step for a designer hoping to put artificial chimps to work in the real world is to translate real world situations into variables the moral module can process. One potential route to achieving this is to design a system to translate real-world information into artificial environment information. So, there would be a pair of systems: the moral module and a real-world translator. The real-world translator translates the real world into a version of the simulated world for the moral module to process.

One possible problem here is that real-world translation will involve interpretative decisions that will prejudice or otherwise problematically bias the moral module. Just as great care must be taken to avoid undermining artificial chimps autonomy; equally great care must be taken to ensure that the real-world translator does not undermine that autonomy.

If all this is successful, the artificial chimp's moral module (complete with real-world translation) could offer responses to moral reasons in the real world. This would be useful in a variety of contexts. Obviously, I have offered only a speculative sketch of the means to do so – but neither the isolation of moral reasons-responsiveness or the translation of relevant real-world variables seem to be impossible, indeed, they seem fairly well within the reach of current technology. In any case, I am speculative largely because there is very little precedence here and everything must start somewhere.

9.6. Artificial Moral Modules and Conceptual Capacities

So far, I have outlined how artificial organisms that meet the historical condition of autonomy and have chimp-level moral reasons-responsiveness could develop. Then, I have offered a sketch of how that moral reasons-responsiveness could be put to work in the real world via a moral module. There remains a pressing question, however. Chimps are not moral agents because they are not *adequately* moral reasons-responsive. Can artificial organisms be adequately moral reasons-responsive like humans are? If so, would they then be moral agents?

Humans are adequately moral reasons-responsive because they have conceptual capacities. If this is a good precedence, then an artificial organism could not be adequately moral reasons-responsive without conceptual capacities. Developing an artificial chimp based moral module with real world translation would be hard. Developing artificial organisms with conceptual capacities and reverse engineering their moral reasons-responsiveness into a moral module would be harder.

Let us take these in turn. First, can artificial organisms develop adequate moral reasons-responsiveness? A designer might hope that evolving conceptual capacities could be reasonably expected from artificial organisms in an artificial life system with sufficient physical and environmental complexity, as evolving biological and psychological capacities can. However, the natural evolution of conceptual capacities seems to be a singularly rare phenomenon. Biological and psychological capacities appear straightforwardly adaptive and have evolved many times, but conceptual capacities have evolved only once. The odds do not favour the designer.

One possibility here is to draw inspiration from the evolutionary history of humanity. Evolutionary theories suggest that humans evolved from chimp-like agents because they faced certain environmental changes. A designer could intervene in the creation or selection of the artificial environment to ensure that these environmental changes regularly occurred. This might, a designer may hope, weight the odds in favour of the development of conceptual capacities in artificial organisms. The increased odds would come at the cost of increasing the causal contribution of the designer's actions. Though this increase seems to me insufficient to make the designer an autonomy-undermining causal source of many of the artificial organisms' actions. In any case, though the odds may be altered, they remain unknown; it is difficult to say whether artificial moral agents would evolve even if certain environmental challenges were guaranteed. After all, other species underwent the same environmental changes without developing conceptual capacities.

A second possibility is to enhance artificial chimps with a 'conceptual capacity' enhancement. Let me explain how this would enable artificial organisms to be adequately moral reasons-responsive without undermining their autonomy. First, I will tackle how this would work, then I will explain why it would not undermine autonomy.

First, a designer must isolate an individual artificial chimp's program. There are then two options, first, a large language model, already trained in human language, can be employed to conceptualise the reasoning process of the artificial chimp, apply inferences and generalisations to that conceptualisation, generating second-order conceptualisations of laws and logic, and feed those back in as new reasons for the artificial chimp. This would 'uplift' the artificial chimp to a hybrid, cyborg-like agent with conceptual capacities. This agent would be a moral agent in its own world, autonomous, responsible, and adequately responding to moral reasons. If it could be suitably integrated into the real world, then it could adequately respond to moral reasons there too. I foresee two concerns about the enhancement's influence on the artificial chimp's autonomy. First, perhaps enhancement in general is autonomy undermining. Second, perhaps the LLM's assumptions about human language are autonomy undermining. Though at first pass there seems a good case for thinking that humans are in a similar situation (i.e., we did not generate our own language and we are happy to enhance ourselves and others (see 7.3. and the cancer resistant human as an example of non-responsibility undermining enhancement).

A third possibility to create a moral module, as described in the previous section, and then implant that module in an artificial agent that is already competent with language. The artificial agent can then interpret the moral module's inputs conceptually, as a kind of 'conscience'. This would be a marriage of ice and fire – the moral module adding (autonomous) moral reasons-responsiveness to an artificial agent that would otherwise lack it. Though this possibility again prompts autonomy-based concerns. Though, crucially, the agent, while it may not be *fully* autonomous, would have autonomous moral reasons-responsiveness.

An artificial agent made by one of these last two method would contain three parts. First, the artificial chimp-based moral module; second, the real-world translator module that can convert real-world information (like images, videos, robotic perceptual data, etc.) to variables that the moral module can process; third, an LLM trained to conceptualise, explain, and feedback into the moral module.

This system would, I believe, overcome all existing arguments against the possibility of artificial moral agents. The system would not (necessarily) be conscious, but I have argued in chapter 2 that consciousness is not a necessary condition of moral agency. The system would be autonomous (which is an important condition of moral agency that many artificial systems fail to meet). It would have psychological capacities for morally reasons-responsiveness and have the capacity to conceptualise and make utterances about its moral reasoning.

This system would does not have the benefit of our significant cultural history and developments. But it *could* learn our culture (in the same way that human children do) or that, while cultural sensitivity *is* practically useful for a moral agent in our own world, it is by no means a necessary condition of moral agency in general.

9.7. Conclusion

I have outlined an artificial system that would have autonomous moral capacities. The development of artificial autonomous moral capacities must be done in a radically different way to traditional artificial system development methods, because all traditional methods undermine the autonomous causal history of the artificial system. I have argued that artificial life systems, systems in which artificial organisms evolve within a digital artificial environment, can, given three conditions, be autonomous. Those three conditions are that the

system has sufficient *environmental complexity*, which means a complex environment that can support different evolutionary strategies, that the system has sufficient *physical complexity*, which means that the artificial organisms can evolve open-endedly, without templates or preset specifications, in small physical increments, and finally, that the system is free from agential intervention from the designers, which I have suggested can be done by randomising and automating environment and 'starter' organism generation.

Autonomous artificial organisms are possible, then, and in section 4 I argued that they are likely able to develop moral reasons-responsiveness. I outlined two variations that could lead to this: the 'psychological strategy' in which artificial organisms initially have representative capacities via (autonomy-conserving) design and subsequently develop psychological capacities and the 'biological strategy' in which artificial organisms independently develop representative capacities and use them to represent the norms implicit in their biological capacities. In either case, the artificial organisms are both autonomous and have psychological capacities for moral reasons-responsiveness (like chimps).

I then discussed some strategies for evolved artificial systems to develop conceptual capacities for moral reasons-responsiveness. Evolving them, either entirely fully autonomously or via environmental intervention, may be a possibility, but should not be counted upon. Two more reliable strategies are to 'uplift' the artificial organisms by integrating their decision-making processes with a machine-learning language model and to 'implant' them into an artificial agent that already has linguistic ability. Doing this while ensuring they are autonomous may be difficult.

If either strategy is successful, the resulting an artificial system ought to be, by almost everyone's lights, an 'artificial moral agent'. It would be responsible (see the end of 8.7.), adequately respond to moral reasons, and hopefully offer some fascinating insights into moral theory (or at least stop rogue artificial systems from wanton harm).

10. Conclusion

10.1. The Bird's Eye View Again

I offer a chapter-based summary below. But first I want to highlight the central conclusions of the thesis.

In Part I, I offered a novel reason for trying to develop advanced artificial systems; I put forward a definition of moral agency agreeable to much of the literature on machine ethics; and I staked out a position within machine ethics by arguing against consciousness conditions, but for intentionality conditions, in moral agency.

In Part II I argued that contemporary artificial systems are not moral agents and cannot be responsible because they have the wrong sort of history, even though they can be adequately morally reasons-responsive. In chapter 6, I developed this general idea with reference to the literature on moral epistemology, while defending the claim that artificial systems can be adequately moral reasons-responsive. In chapter 7, I developed it with reference to the literature on responsibility, developing a historical condition of autonomy which can accommodate design cases and arguing that some designed agents can be autonomous, but contemporary artificial agents are not.

In Part III, I outlined theories for how morality evolved and argued that artificial organisms can evolve to be moral. In chapter 8, I sketched theories of how morality evolved, offering a categorisation of evolved capacities for moral reasons-responsiveness that can be used as a developmental template. In chapter 9, I argued that artificial organisms can be both autonomous and adequately moral reasons-responsive, and therefore moral agents.

10.2. Points of Disengagement

The thesis covers a lot of ground, and there are many details that remain to be filled in by further research. The reader may not be fully convinced by each and every argument I have advanced. I take this opportunity to highlight means for a skeptical reader to find value despite potential disagreement. That is, there may be some occasions where a reader may disengage with my arguments, but I want to point out that this disengagement can (with a little interpretative work) be temporary rather than terminal.

If you find yourselves disagreeing with me about the precise conditions of moral agency, you may yet find that my arguments about reliance on deference and historical conditions for autonomy have a place in your preferred conditions of moral agency. Most of those who prefer a consciousness condition face an additional challenge in the Moral Decision Machine – which can respond to moral reasons adequately without being conscious; but they may still endorse anti-deference and historical conditions for responsibility. If you can accommodate those conditions, then you have reason to be interested in the evolutionary methodology of later chapters too.

If you are unconvinced that only non-agent forces can be autonomy-conserving external causal sources, that autonomy (or responsibility) has historical conditions, or that deference poses a problem for moral agency, then you probably think that there are other means of designing autonomous agents and may question the focus on evolution. It is true that I focus on evolutionary forces because they seem to me the *best and only* way of conserving-autonomy in designed agents, but I think that focus can be justified even if they are not. Evolutionary forces are valuable not just because they are autonomy-conserving, but because they are also viable routes to moral and cognitive diversity. Even if it is *possible* to 'game the system', and design autonomous agents that nonetheless do the bidding of humans, derive their functioning from the conceptual imaginings of a human designer, or substantially mimic humans, then they would not offer the diverse standpoint that I hope for. The discussion of evolving artificial agents in chapter 9, then, is not valuable only because it is the sole route to autonomous artificial agents. Evolutionary artificial agents offer kinds of value that other types of artificial agent may not and are worth pursuing on that basis anyway. Putting other types of value aside, even if evolutionary forces are but one autonomy-conserving means of design among many then those alternatives can be understood to lie beyond my scope, and my conclusions can carry the disclaimer that other autonomy-conserving design methodologies may also produce moral agents.

Finally, and probably most importantly, you may be unconvinced that artificial organisms could develop conceptual capacities and believe that both the strategies for doing so would fail. If so, I would point to the value of evolving artificial moral reasons-responsiveness *at all*. Even if you remain unconvinced that artificial moral agency is possible, there is a stronger case for the possibility of animal-like (or chimp-like) artificial organisms, and this is worth considering on its own. Especially if their capacities could be put to work in the kind of 'moral module' I suggested. Even a negative conclusion, though I have argued against it, that

no artificial organism could evolve to be moral would be substantial progress, if only for ruling *out* evolutionary development based methodologies for artificial moral agency.

10.3. Chapter Summaries

In the introduction I put forward my position that philosophy, and philosophers, aim for objectivity. To get closer to objectivity, I said, humans need a non-human perspective. In moral philosophy, that means non-human moral agents. Artificial moral agents are the non-human moral agents on the horizon – I wanted to see whether they were a mirage. That is, I wanted to learn whether artificial moral agents are possible.

To help motivate the question, Chapter 2 laid out a pro tanto reason for why we should want advanced artificial systems in the first place. Advanced artificial systems, I argued, would increase cognitive diversity; and greater cognitive diversity offers epistemic advantages for our group's theories. Artificial moral agents would be an example of advanced artificial systems and offer epistemic advantages for our theories by contributing to cognitive diversity. They may also offer an analogous advantage for our moral theories. Artificial moral agents could adopt diverse moral perspectives that offer an epistemic advantage for our group's moral theories.

Motivations thus laid out, I proceeded to the central thesis of the thesis. Broadly, my position is that artificial systems can be moral agents, but only under certain conditions. To begin the defence of this, the next three chapters of the thesis aimed to establish a solid foundation for the concept of 'moral agency'.

First on the agenda, in chapter 3, were some starting conditions. First, adult humans are the paradigmatic moral agents. Adult humans were the exemplar and guideline in the discussion of moral agency that followed. Second, while humans are the prototype moral agents, moral agency does not necessarily describe exclusively humans. On the contrary, I assumed that it is necessary that some non-human *could* satisfy the conditions of moral agency. Then, I set out two necessary conditions for moral agency: to be a moral agent, an agent must be both adequately moral reasons-responsive and responsible. I defended these conditions as conceptually necessary for moral agency.

Second up were two debated potential conditions for moral agency. Consciousness and intentionality. In chapter 4, I argued that consciousness was not a condition of moral agency,

because, chiefly, the concept of consciousness is vague. Perhaps consciousness is epistemically inaccessible, as functionalists argue, and in which case it is a poor condition for moral agency. Perhaps consciousness is epistemically accessible, but if so, I argued, only humans can be conscious. I acknowledged that humans are the prototype moral agents, but the consciousness condition entailed that *only* humans could be moral agents. Prototype or not, this seemed unpalatable, since I assume that it must be possible in principle for a non-human to be moral agent. So, I rejected the consciousness condition for moral agency.

In chapter 5 I defended an intentionality condition for moral agency. The reasoning here was quite different to that for the consciousness condition. One very plausible theory of intentional states is that they are functional states. I argued that any reasonable functionalist, which at this point contains all those who want to entertain the possibility of non-human moral agents, ought to adopt functionalism about intentional states. If they do, I argued, then they may as well include an intentionality condition in moral agency, since everyone agrees that all moral agents must have functionalist intentional states.

The end of chapter 5 was the end of Part I. The concept of moral agency defended was that an agent was a moral agent if they were adequately moral reasons-responsive and responsible. I argued that to satisfy these conditions moral agents must have (functionalist) intentional, but not necessarily conscious, states. Having intentional states is not a problem for artificial agents, and part I found no convincing arguments for denying that artificial agents can be moral agents.

Part II considered further potential conditions of moral agency: avoiding deference, and autonomy. I argued that potential conditions about avoiding deference and autonomy justify concerns about the possibility of artificial moral agency. But, I argued, these concerns could be answered, and the possibility of artificial moral agents defended.

Chapter 6 argued for two conclusions. First, artificial agents can be adequately moral reasons-responsive, and thus satisfy one of the two necessary conditions for moral agency. I argued for this with the thought experiment of 'The Moral Decision Machine'. The Moral Decision Machine wields a vast database of human decisions and can act upon a decision that perfectly matches any situation it faces. It thus responds precisely to the same moral reasons that the human did, and, in this, demonstrates adequate moral reasons-responsiveness. However, the second conclusion of Chapter 6 is that this was insufficient for moral agency, because the Moral Decision Machine was not responsible for its actions, as it always deferred

to the human (which is generally accepted to be incompatible with responsibility). I argued that all contemporary artificial systems seem to exhibit a similar kind of deference, and thus that none of them were likely to be responsible.

Chapter 7 focused in more detail on why artificial systems seem incapable of responsibility. I considered the design hypothesis: that designed agents could not be autonomous. I discussed the relevant cases from the responsibility literature – 'manipulation' and 'design' cases; and adopted a specific view of the historical conditions for autonomy where an autonomous agent has no other agent as the causal source of (most of) their actions. According to this condition, it turned out that the design hypothesis was false: *some* designed agents *could* be autonomous. However, contemporary artificial systems do have their autonomy undermined by their designers and therefore cannot be autonomous or responsible. To design autonomous agents, I concluded, a designer needed to avoid being the causal source of the designed agent's actions. The foremost way to do so, I suggested, was to design evolutionary agents, since evolutionary forces offer an alternative and autonomy-conserving causal source. Finally, I argued that indeterministic free will is not a necessary condition for responsibility or moral agency.

Part II's central conclusion was that to be moral agent, an agent needs the right kind of history. Furthermore, an adequately morally reasons-responsive agent with the right kind of history is likely to be a moral agent. Evolution offers the right kind of history for autonomy, so Part III turned to the evolutionary method of designing artificial agents.

Chapter 8 discussed the evolution of human moral reasons-responsiveness. I divided capacities for moral reasons-responsiveness into three: Biological, psychological and conceptual capacities. Humans have all three capacities, though the foremost human capacity to respond to moral reasons uses psychological capacities and conceptual capacities in tandem. I concluded that the three capacities as humans possess them are jointly sufficient for adequate moral reasons-responsiveness, and that they could in principle evolve in any agent that followed a similar evolutionary pathway.

Chapter 9 argued that artificial organisms, which together with their simulated environment constitute artificial life systems, can be designed to be autonomous. I argued that based on how moral reasons-responsiveness evolved in the natural world, autonomous artificial organisms could be expected to reach chimp-level moral reasons-responsiveness (i.e., possess biological and psychological, but not conceptual, capacities for moral reasons-

responsiveness). However, conceptual capacities, and thus adequate moral reasons-responsiveness, are less likely to evolve. I suggested two strategies for endowing artificial organisms with conceptual capacities: enhancement through LLMs and designing artificial environments with features that seemed instrumental in prompting the human evolution of conceptual capacities. Should either strategy be successful, which seems at the very least possible, then the artificial organisms would be artificial moral agents. Thus, artificial moral agents are possible, but, crucially, they must be designed such that they have the right kind of history as well as being adequately reasons-responsive, and the only means of achieving that, as far as I can see, is by evolving them in the manner described in this chapter. Chapter 9 also discussed possibilities for using artificial organisms with moral reasons-responsiveness in the real world – something that would be of value even if they are not fully fledged moral agents.

11. Reference List

Aggarwal, I., & Woolley, A. (2018). Team Creativity, Cognition, and Cognitive Style Diversity. *Management Science*, *65*. https://doi.org/10.1287/mnsc.2017.3001

Aguilar, W., Santamaría-Bonfil, G., Froese, T., & Gershenson, C. (2014). The Past, Present, and Future of Artificial Life. *Frontiers in Robotics and AI*, *1*. https://www.frontiersin.org/articles/10.3389/frobt.2014.00008

Alexander, R. D. (1987). *The biology of moral systems.* Aldine de Gruyter.

Allchin, D. (2009). The Evolution of Morality. *Evolution: Education and Outreach*, *2*(4), Article 4. https://doi.org/10.1007/s12052-009-0167-7

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, Wallach, W., & Smit, I. (2006). Why Machine Ethics? *IEEE Intelligent Systems*, *21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83

Anderson, E. (2006). The Epistemology of Democracy. *Episteme*, *3*(1–2), 8–22. https://doi.org/10.3366/epi.2006.3.1-2.8

Anderson, S. L., & Anderson, M. (2020). AI and ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-020-00003-6

Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/

Ayer, A. J. (1963). Freedom and Necessity. In A. J. Ayer, *Philosophical Essays*. Springer.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Beckman, S., Lau, K., & Agogino, A. (2012). Diversity in Design Teams: An Investigation of Learning Styles and their Impact on Team Performance and Innovation. *International Journal of Engineering Education*, *28*.

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, *30*(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8

Bekoff, M., & Pierce, J. (2010). *Wild Justice: The Moral Lives of Animals*. University of Chicago Press.

Berlekamp, E. R., Conway, J. H., & Guy, R. K. (2003). *Winning Ways for Your Mathematical Plays, Vol. 2* (2nd edition). A K Peters/CRC Press.

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, *2*(16). https://doi.org/10.51291/2377-7478.1200

Birch, J. (2022). Materialism and the Moral Status of Animals. *Philosophical Quarterly*, *72*(4), 795–815. https://doi.org/10.1093/pq/pqab072

Björnsson, G. (2016). Outsourcing the deep self: Deep self discordance does not explain away intuitions in manipulation arguments. *Philosophical Psychology*, *29*(5), 637–653. https://doi.org/10.1080/09515089.2016.1150448

Björnsson, G., & Hess, K. (2017). Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents. *Philosophy and Phenomenological Research*, *94*(2), 273–298. https://doi.org/10.1111/phpr.12260

BonJour, L. (1988). *The Structure of Empirical Knowledge* (2nd printing edition). Harvard University Press.

Borgo, S., Franssen, M., Garbacz, P., Kitamura, Y., Mizoguchi, R., & Vermaas, P. E. (2014). Technical artifacts: An integrated perspective. *Applied Ontology* 9 217–235 https://doi.org/10.3233/AO-140137

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bratman, M. E. (1997). Responsibility and Planning. *The Journal of Ethics*, *1*(1), 27–43. https://doi.org/10.1023/A:1009703818699

Bringsjord, S. (2008). Ethical Robots: The Future Can Heed Us. *AI and Society*, *22*(4), 539–550. https://doi.org/10.1007/s00146-007-0090-9

Brooks, A. W. (2014). Get excited: Reappraising pre-performance anxiety as excitement. *Journal of Experimental Psychology. General*, *143*(3), 1144–1158. https://doi.org/10.1037/a0035325

Bryson, J. J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (pp. 63–74).

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26. https://doi.org/10.1007/s10676-018-9448-6

Carruthers, P. (2019). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press.

Chalmers, D. J. (2006). Phenomenal Concepts and the Explanatory Gap. In *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism*. Oxford University Press.

Chalmers, D. J. (2018). The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, *25*(9–10), 6–61.

Champagne, M. (2021). The Mandatory Ontology of Robot Responsibility. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 448–454. https://doi.org/10.1017/S0963180120000997

Chomanski, B. (2019). What's Wrong with Designing People to Serve? *Ethical Theory and Moral Practice*, *22*(4), 993–1015.

Chudnoff, E. (2016). *Intuition* (Reprint edition). Oxford University Press.

Churchland, P. (2021). *Conscience: The Origins of Moral Intuition* (Reprint edition). W. W. Norton & Company.

Clarridge, A. (2009). *Cellular Automata: Algorithms and Applications* [Msc]. Queen's University.

Clement, G. (2013). Animals and Moral Agency: The Recent Debate and Its Implications. *Journal of Animal Ethics*, *3*(1), 1–14. https://doi.org/10.5406/janimalethics.3.1.0001

Clune, J., Goldsby, H. J., Ofria, C., & Pennock, R. T. (2011). Selective pressures for accurate altruism targeting: Evidence from digital evolution for difficult-to-test aspects of inclusive fitness theory. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1706), 666–674. https://doi.org/10.1098/rspb.2010.1557

Cochrane, A. (2018). *Sentientist Politics: A Theory of Global Inter-Species Justice*. Oxford University Press.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & SOCIETY*, *24*(2), 181–189. https://doi.org/10.1007/s00146-009-0208-3

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. Palgrave-Macmillan.

Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, *27*(1), 61–77. https://doi.org/10.1007/s13347-013-0133-8

Collins, S. (2023). Group blameworthiness and group rights. *Inquiry*. https://doi.org/10.1080/0020174X.2023.2191651

Crisp, R. (2018). Prudential and Moral Reasons. In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199657889.013.0035

Cruz, H. D., & Smedt, J. D. (2013). *The value of epistemic disagreement in scientific practice. The case of Homo floresiensis*. https://philarchive.org/rec/CRUTVO

Curry, O. S. (2016). Morality as Cooperation: A Problem-Centred Approach. In T. K. Shackelford & R. D. Hansen (Eds.), *The Evolution of Morality* (pp. 27–51). Springer International Publishing. https://doi.org/10.1007/978-3-319-19671-8_2

Cyr, T. W. (2019). Moral Responsibility, Luck, and Compatibilism. *Erkenntnis*, *84*(1), 193–214. https://doi.org/10.1007/s10670-017-9954-7

Cyr, T. W. (2020). Manipulation and constitutive luck. *Philosophical Studies*, *177*(8), 2381–2394. https://doi.org/10.1007/s11098-019-01315-y

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

De Marco, G. (2023). Manipulation, machine induction, and bypassing. *Philosophical Studies*, *180*(2), 487–507. https://doi.org/10.1007/s11098-022-01906-2

Deery, O., & Nahmias, E. (2017). Defeating Manipulation Arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*, *174*(5), 1255–1276. https://doi.org/10.1007/s11098-016-0754-8

DeGrazia, D. (1996). *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge University Press.

DeGrazia, D. (2020). Sentience and Consciousness as Bases for Attributing Interests and Moral Status: Considering the Evidence and Speculating Slightly Beyond. In L. S. M. Johnson, A. Fenton, & A. Shriver (Eds.), *Neuroethics and Nonhuman Animals* (pp.

17–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-31011-0_2

Dennett, D. C. (1981). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

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. https://doi.org/10.1080/09528130110100289

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Dretske, F. I. (1973). Perception and Other Minds. *Noûs*, *7*(1), 34–44. https://doi.org/10.2307/2216182

Dreyfus, H. L. (1992). *What Computers Still Can?T Do: A Critique of Artificial Reason*. MIT Press.

Dunbar, R. I. M. (1996). *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.

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. https://doi.org/10.1007/s11948-022-00408-y

Eiben, A. E. (2014). Grand Challenges for Evolutionary Robotics. *Frontiers in Robotics and AI*, *1*. https://www.frontiersin.org/articles/10.3389/frobt.2014.00004

Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-Year Olds, but Not Chimpanzees, Attempt to Manage Their Reputations. *PLOS ONE*, *7*(10), e48433. https://doi.org/10.1371/journal.pone.0048433

Engelmann, J. M., & Tomasello, M. (2018). The middle step: Joint intentionality as a human-unique form of second-personal engagement. In *The Routledge handbook of collective intentionality* (pp. 433–446). Routledge/Taylor & Francis Group.

Fischer, J. M. (2004). Responsibility and Manipulation. *The Journal of Ethics*, *8*(2), 145–177. https://doi.org/10.1023/B:JOET.0000018773.97209.84

Fischer, J. M. (2011). The Zygote Argument remixed. *Analysis*, *71*(2), 267–272.

Fischer, J. M. (2014). Review of Free will, agency, and meaning in life, by Derk Pereboom. *Science, Religion, and Culture*, *1*, 202–208.

Fischer, J. M. (2016). How Do Manipulation Arguments Work? *The Journal of Ethics*, *20*(1), 47–67. https://doi.org/10.1007/s10892-016-9225-x

Fischer, J. M. (2017). Responsibility, Autonomy, and the Zygote Argument. *The Journal of Ethics*, *21*(3), 223–237.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility* (M. Ravizza, Ed.). New York: Cambridge University Press.

FitzPatrick, W. (2021). Morality and Evolutionary Biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/morality-biology/

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/b:mind.0000035461.63578.9d

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Frankfurt, H. G. (1988). *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, H. G. (2002). Reply to Fischer. In *Contours of Agency: Essays on Themes from Harry Frankfurt Edited by Sarah Buss and Lee Overton*. MIT Press.

Frankish, K. (Ed.). (2017). *Illusionism: As a theory of consciousness* (1st edition). Imprint Academic.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Garwood, R. J., Spencer, A. R. T., & Sutton, M. D. (2019). REvoSim: Organism-level simulation of macro and microevolution. *Palaeontology*, *62*(3), 339–355. https://doi.org/10.1111/pala.12420

Gibbard, A. (1992). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Clarendon Press.

Gilby, I. C., & Machanda, Z. P. (2022). Advanced cognition in wild chimpanzees: Lessons from observational studies. *Current Opinion in Behavioral Sciences*, *46*, 101183. https://doi.org/10.1016/j.cobeha.2022.101183

Godfrey-Smith, P. (2018). *Other Minds: The Octopus and the Evolution of Intelligent Life* (Reprint edition). William Collins.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the Moral Domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77. https://doi.org/10.1016/j.cognition.2017.03.004

Grim, P., Singer, D. J., Bramson, A., Holman, B., McGeehan, S., & Berger, W. J. (2019). Diversity, Ability, and Expertise in Epistemic Communities. *Philosophy of Science*, *86*(1), 98–123. https://doi.org/10.1086/701070

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology*, *10*(2), 115–121. https://doi.org/10.1007/s10676-008-9163-9

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on Ai, Robots, and Ethics*. MIT Press.

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

Habermas, J. (2003). *The Future of Human Nature*. Polity.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295x.108.4.814

Haji, I. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*.

Hakli, R., & Mäkelä, P. (2016). Robots, Autonomy, and Responsibility. In J. Seibt, M. Nørskov, & S. S. Andersen (Eds.), *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016* (pp. 145–154). Amsterdam, The Netherlands: IOS Press.

Hakli, R., & Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist*, *102*(2), 259–275. https://doi.org/10.1093/monist/onz009

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Hanczyc, M. M. (2020). Engineering life: A review of synthetic biology. *Artificial Life*, *26*(2), 260–273.

Harding, S. (2016). *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Cornell University Press. https://muse.jhu.edu/pub/255/monograph/book/48914

Haugeland, J. (1990). The Intentionality All-Stars. *Philosophical Perspectives*, *4*, 383–427. https://doi.org/10.2307/2214199

Heil, J. (2019). *Philosophy of Mind: A Contemporary Introduction*.

Hellström, T. (2013). On the Moral Responsibility of Military Robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Herdova, M. (2021). The importance of being Ernie. *Thought: A Journal of Philosophy*, *10*(4), 257–263. https://doi.org/10.1002/tht3.503

Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics*, *120*(1), 94–127. JSTOR. https://doi.org/10.1086/648610

Hills, A. (2020). Moral Testimony: Transmission Versus Propagation. *Philosophy and Phenomenological Research*, *101*(2), 399–414. https://doi.org/10.1111/phpr.12595

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. https://doi.org/10.1007/s10676-008-9167-5

Hoever, I. J., van Knippenberg, D., van Ginkel, W. P., & Barkema, H. G. (2012). Fostering team creativity: Perspective taking as key to unlocking diversity's potential. *Journal of Applied Psychology*, *97*, 982–996. https://doi.org/10.1037/a0029159

Holman, B., Berger, W. J., Singer, D. J., Grim, P., & Bramson, A. (2018). Diversity and Democracy: Agent-Based Modeling in Political Philosophy. *Historical Social Research / Historische Sozialforschung*, *43*(1 (163)), 259–284.

Horwitz, S. K., & Horwitz, I. B. (2007). The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography. *Journal of Management*, *33*(6), 987–1015. https://doi.org/10.1177/0149206307308587

Howell, R. J. (2014). Google Morals, Virtue, and the Asymmetry of Deference. *Noûs*, *48*(3), 389–415. https://doi.org/10.1111/j.1468-0068.2012.00873.x

Huemer, M. (2005). *Ethical Intuitionism*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230597051

Hutto, D. D., & Satne, G. (2015). The Natural Origins of Content. *Philosophia*, *43*(3), 521–536. https://doi.org/10.1007/s11406-015-9644-0

Hyslop, A. (2019). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/other-minds/

Hyslop, A., & Jackson, F. (1972). The Analogical Inference to Other Minds. *American Philosophical Quarterly*, *9*(3), 168–176.

Intemann, K. (2009). Why Diversity Matters: Understanding and Applying the Diversity Component of the National Science Foundation's Broader Impacts Criterion. *Social Epistemology*, *23*(3–4), 249–266. https://doi.org/10.1080/02691720903364134

Intemann, K. (2010). 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia*, *25*(4), 778–796.

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, *5*(2), 217–232. https://doi.org/10.1007/s13194-014-0105-6

Ioan, M., & Howard, D. (2017). Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. In T. Powers (Ed.), *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*. Springer.

Jackson, F., & Pettit, P. (1990). In defence of folk psychology. *Philosophical Studies*, *59*(1), 31–54. https://doi.org/10.1007/BF00368390

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5

Johnson, L. S. M. (2021). Moral Status and the Consciousness Criterion. In L. S. M. Johnson & L. S. M. Johnson (Eds.), *The Ethics of Uncertainty: Entangled Ethical and*

Epistemic Risks in Disorders of Consciousness (p. 0). Oxford University Press. https://doi.org/10.1093/med/9780190943646.003.0007

Joyce, R. (2006). *The Evolution of Morality*. MIT Press.

Kagan, S. (2019). *How to Count Animals, More Or Less*. Oxford University Press.

Kane, R. (1998). *The Significance of Free Will*. Oxford University Press.

Kauppinen, A. (2013). A Humean theory of moral intuition. *Canadian Journal of Philosophy*, *43*(3), 360–381. https://doi.org/10.1080/00455091.2013.857136

Kauppinen, A. (2022). Moral Sentimentalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/moral-sentimentalism/

Kim, K.-J., & Cho, S.-B. (2006). A Comprehensive Overview of the Applications of Artificial Life. *Artificial Life*, *12*(1), 153–182. https://doi.org/10.1162/106454606775186455

King, M. (2013). The Problem with Manipulation. *Ethics*, *124*(1), 65–83. https://doi.org/10.1086/671391

Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.

Korsgaard, C. M. (2006). *Morality and the Distinctiveness of Human Action*. https://doi.org/10.1515/9781400830336-008

Kriegel, U. (2019). The Value of Consciousness. *Analysis*, *79*(3), 503–520. https://doi.org/10.1093/analys/anz045

Landes, J. (2020). Variety of Evidence. *Erkenntnis*, *85*(1), 183–223. https://doi.org/10.1007/s10670-018-0024-6

Laukyte, M. (2014). Artificial Agents: Some Consequences of a Few Capacities. In *Sociable Robots and the Future of Social Relations* (pp. 115–122). IOS Press. https://doi.org/10.3233/978-1-61499-480-0-115

Laukyte, M. (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology*, *19*(1), 1–17. https://doi.org/10.1007/s10676-016-9411-3

Levy, N. (2009). Luck and History-Sensitive Compatibilism. *The Philosophical Quarterly*, *59*(235), 237–251. https://doi.org/10.1111/j.1467-9213.2008.568.x

List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology*, *34*(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford, GB: Oxford University Press.

Liu, X. (2022). Manipulation and Machine Induction. *Mind*, *131*(522), 535–548. https://doi.org/10.1093/mind/fzaa085

Longino, H. E. (2018). The Fate of Knowledge. In *The Fate of Knowledge*. Princeton University Press. https://doi.org/10.1515/9780691187013

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McGrath, S. (2009). THE PUZZLE OF PURE MORAL DEFERENCE. *Philosophical Perspectives*, *23*, 321–344. JSTOR.

McKenna, M. (2004). Responsibility and Globally Manipulated Agents. *Philosophical Topics*, *32*(1/2), 169–192.

McKenna, M. (2012). Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism. *The Journal of Ethics*, *16*(2), 145–174.

McKenna, M. (2016). A Modest Historical Theory of Moral Responsibility. *The Journal of Ethics*, *20*(1/3), 83–105.

McLaughlin, P. (2000). *What Functions Explain: Functional Explanation and Self-Reproducing Systems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498510

McPherson, T. (2022). Supervenience in Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/supervenience-ethics/

Mele, A. R. (1995). *Autonomous Agents: From Self Control to Autonomy*. New York, US: Oxford University Press.

Mele, A. R. (2008). *Free Will and Luck* (1st edition). Oxford University Press.

Mele, A. R. (2013). Manipulation, Moral Responsibility, and Bullet Biting. *The Journal of Ethics*, *17*(3), 167–184. https://doi.org/10.1007/s10892-013-9147-9

Mele, A. R. (2020). Moral responsibility and manipulation: On a novel argument against historicism. *Philosophical Studies*, *177*(10), 3143–3154. https://doi.org/10.1007/s11098-019-01363-4

Melis, A. P., Hare, B., & Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science (New York, N.Y.)*, *311*(5765), 1297–1300. https://doi.org/10.1126/science.1123007

Mello, A. L., & Delise, L. A. (2015). Cognitive Diversity to Team Outcomes: The Roles of Cohesion and Conflict Management. *Small Group Research*, *46*(2), 204–226. https://doi.org/10.1177/1046496415570916

Mello, A. L., & Rentsch, J. R. (2015). Cognitive diversity in teams: A multidisciplinary review. *Small Group Research*, *46*, 623–658. https://doi.org/10.1177/1046496415602558

Melnyk, A. (1994). Inference to the Best Explanation and Other Minds. *Australasian Journal of Philosophy*, *72*(4), 482–491. https://doi.org/10.1080/00048409412346281

Miller, C. C., Burke, L. M., & Glick, W. H. (1998). Cognitive Diversity among Upper-Echelon Executives: Implications for Strategic Decision Processes. *Strategic Management Journal*, *19*(1), 39–58.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R. G. (1999). Wings, Spoons, Pills, and Quills: A Pluralist Theory of Function. *The Journal of Philosophy*, *96*(4), 191–206. https://doi.org/10.2307/2564702

Milliken, F. J., & Martins, L. L. (1996). Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups. *Academy of Management Review*, *21*(2), 402–433. https://doi.org/10.5465/amr.1996.9605060217

Mogensen, A. L. (2017). Moral Testimony Pessimism and the Uncertain Value of Authenticity. *Philosophy and Phenomenological Research*, *95*(2), 261–284. https://doi.org/10.1111/phpr.12255

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80

Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. *AI & SOCIETY*, *36*(2), 429–443. https://doi.org/10.1007/s00146-020-01002-1

Muldoon, R. (2013). Diversity and the Division of Cognitive Labor. *Philosophy Compass*, *8*(2), 117–125. https://doi.org/10.1111/phc3.12000

Muntean, I., & Howard, D. (2016). A Minimalist Model of the Artificial Autonomous Moral Agent (Aama). In *SSS-16 Symposium Technical Reports. Association for the Advancement of Artificial Intelligence*. AAAI.

Musiał, M. (2022). Can we design artificial persons without being manipulative? *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01575-z

Nagel, T. (1989). *The View from Nowhere*. Oxford University Press.

Neander, K. (2017). *A Mark of the Mental: A Defence of Informational Teleosemantics*. Cambridge, USA: MIT Press.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds* (S. P. Stich, Ed.). Oxford, GB: Oxford University Press.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51–62. https://doi.org/10.1007/s10676-013-9335-0

Nowak, M. A., McAvoy, A., Allen, B., & Wilson, E. O. (2017). The general form of Hamilton's rule makes no predictions and cannot be tested empirically. *Proceedings of the National Academy of Sciences*, *114*(22), 5665–5670. https://doi.org/10.1073/pnas.1701805114

Ofria, C., & Wilke, C. O. (2004). Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life*, *10*(2), 191–229. https://doi.org/10.1162/106454604773563612

Ofria, C., & Wilke, C. O. (2005). Avida: Evolution Experiments with Self-Replicating Computer Programs. *Artificial Life Models in Software*, 3–35. https://doi.org/10.1007/1-84628-214-4_1

Olson, B. J., Parayitam, S., & Bao, Y. (2007). Strategic Decision Making: The Effects of Cognitive Diversity, Conflict, and Trust on Decision Outcomes. *Journal of Management*, *33*(2), 196–222. https://doi.org/10.1177/0149206306298657

Olsson, E. (2021). Coherentist Theories of Epistemic Justification. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab,

Stanford University. https://plato.stanford.edu/archives/fall2021/entries/justep-coherence/

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity* (Illustrated edition). Hachette Books.

Osimani, B., & Landes, J. (2023). Varieties of Error and Varieties of Evidence in Scientific Inference. *The British Journal for the Philosophy of Science*, *74*(1), 117–170. https://doi.org/10.1086/714803

Owe, A., Baum, S. D., & Coeckelbergh, M. (2022). Nonhuman Value: A Survey of the Intrinsic Valuation of Natural and Artificial Nonhuman Entities. *Science and Engineering Ethics*, *28*(5), 38. https://doi.org/10.1007/s11948-022-00388-z

Page, S. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)*. Princeton University Press. https://doi.org/10.2307/j.ctt7sp9c

Papineau, D. (2002). *Thinking About Consciousness*. Oxford, GB: Oxford University Press UK.

Pargetter, R. (1984). The Scientific Inference to Other Minds. *Australasian Journal of Philosophy*, *62*(2), 158–163. https://doi.org/10.1080/00048408412341341

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, *05*(02), 105–129. https://doi.org/10.1142/S1793843013500017

Pereboom, D. (2001). *Living without Free Will*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498824

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. OUP Oxford.

Peters, U., Honeycutt, N., Block, A. D., & Jussim, L. (2020). Ideological Diversity, Hostility, and Discrimination in Philosophy. *Philosophical Psychology*, *33*(4), 511–548. https://doi.org/10.1080/09515089.2020.1743257

Pieterse, A. N., van Knippenberg, D., & van Ginkel, W. P. (2011). Diversity in goal orientation, team reflexivity, and team performance. *Organizational Behavior and Human Decision Processes*, *114*, 153–164. https://doi.org/10.1016/j.obhdp.2010.11.003

Powers, T. M. (2013). On the Moral Agency of Computers. *Topoi*, *32*(2), 227–236. https://doi.org/10.1007/s11245-012-9149-4

Preston, B. (2009). Philosophical Theories of Artifact Function. In A. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences* (pp. 213–233). North-Holland. https://doi.org/10.1016/B978-0-444-51667-1.50013-6

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y

Pust, J. (2016). *Intuitions as Evidence* (1st edition). Routledge.

Putnam, H. (1975). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

Queller, D. C. (1992). A General Model for Kin Selection. *Evolution*, *46*(2), 376–380. https://doi.org/10.1111/j.1558-5646.1992.tb02045.x

Queller, D. C., & Strassmann, J. E. (1998). Kin Selection and Social Insects. *BioScience*, *48*(3), 165–175. JSTOR. https://doi.org/10.2307/1313262

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA, USA: MIT Press.

Ratnieks, F. (1988). Reproductive Harmony via Mutual Policing by Workers in Eusocial Hymenoptera. *The American Naturalist*, *132*(2), 217–236. https://doi.org/10.1086/284846

Ratnieks, F., & Wenseleers, T. (2008). Altruism in insect societies and beyond: Voluntary or enforced? *Trends in Ecology & Evolution*, *23*(1), 45–52. https://doi.org/10.1016/j.tree.2007.09.013

Ray, T. S. (1993). An Evolutionary Approach to Synthetic Biology: Zen and the Art of Creating Life. *Artificial Life*, *1*(1_2), 179–209. https://doi.org/10.1162/artl.1993.1.1_2.179

Rottschaefer, W. A. (1998). *The Biology and Psychology of Moral Agency*. Cambridge University Press.

Rowlands, M. (2012). *Can Animals Be Moral?* Oxford University Press.

Sapontzis, S. F. (1992). *Morals, Reason, and Animals*. Temple University Press.

Sarkar, S. (2010). Diversity: A Philosophical Perspective. *Diversity*, *2*(1), Article 1. https://doi.org/10.3390/d2010127

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

Schlosser, M. (2019). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

Schmidt, M. F. H., & Tomasello, M. (2016). How chimpanzees cooperate: If dominance is artificially constrained. *Proceedings of the National Academy of Sciences*, *113*(44), E6728–E6729. https://doi.org/10.1073/pnas.1614378113

Schulzke, M. (2013). Autonomous Weapons and Distributed Responsibility. *Philosophy and Technology*, *26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies In Philosophy*, *39*(1), 98–119. https://doi.org/10.1111/misp.12032

Shapiro, P. (2006). Moral agency in other animals. *Theoretical Medicine and Bioethics*, *27*(4), 357–373. https://doi.org/10.1007/s11017-006-9010-0

Shepherd, J. (2023). Non-Human Moral Status: Problems with Phenomenal Consciousness. *AJOB Neuroscience*, *14*(2), 148–157. https://doi.org/10.1080/21507740.2022.2148770

Shevlin, H. (2021). How Could We Know When a Robot was a Moral Patient? *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 459–471. https://doi.org/10.1017/S0963180120001012

Shoemaker, D. (2003). Caring, Identification, and Agency. *Ethics*, *114*(1), 88–118. https://doi.org/10.1086/376718

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.

Silk, J. B. (2007). Social components of fitness in primate groups. *Science (New York, N.Y.)*, *317*(5843), 1347–1351. https://doi.org/10.1126/science.1140734

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Sinnott-Armstrong, W., & Conitzer, V. (2021). *How Much Moral Status Could Artificial Intelligence Ever Achieve?* (pp. 269–289). https://doi.org/10.1093/oso/9780192894076.003.0016

Sinnott-Armstrong, W., & Miller, C. B. (Eds.). (2007). Symbolic Thought and the Evolution of Human Morality. In *Moral Psychology*. The MIT Press. https://doi.org/10.7551/mitpress/7481.003.0007

Sison, A. J. G., & Redín, D. M. (2023). A neo-aristotelian perspective on the need for artificial moral agents (AMAs). *AI & SOCIETY*, *38*(1), 47–65. https://doi.org/10.1007/s00146-021-01283-0

Sliwa, P. (2012). In defense of moral testimony. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *158*(2), 175–195. JSTOR.

Sober, E., & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Solomon, M. (2001). *Social Empiricism*. Cambridge, MA, USA: MIT Press.

Solomon, M. (2006). Norms of Epistemic Diversity. *Episteme*, *3*(1–2), 23–36. https://doi.org/10.3366/epi.2006.3.1-2.23

Søvik, A. O. (2022). How a non-conscious robot could be an agent with capacity for morally responsible behaviour. *AI and Ethics*, *2*(4), 789–800. https://doi.org/10.1007/s43681-022-00140-0

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow, R. (2021). Why machines cannot be moral. *AI & SOCIETY*. https://doi.org/10.1007/s00146-020-01132-6

Steel, D., Fazelpour, S., Crewe, B., & Gillette, K. (2021). Information elaboration and epistemic effects of diversity. *Synthese*, *198*(2), 1287–1307. https://doi.org/10.1007/s11229-019-02108-w

Steels, L., & Brooks, R. (2018). *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Routledge.

Sterelny, K. (2003). *Thought In A Hostile World: The Evolution of Human Cognition*. WB.

Sterelny, K. (2014). *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *127*(1), 109–166.

Suchak, M., Eppley, T. M., Campbell, M. W., Feldman, R. A., Quarles, L. F., & de Waal, F. B. M. (2016). How chimpanzees cooperate in a competitive world. *Proceedings of the National Academy of Sciences*, *113*(36), 10215–10220. https://doi.org/10.1073/pnas.1611826113

Suchak, M., & Waal, F. B. M. de. (2016). Reply to Schmidt and Tomasello: Chimpanzees as natural team-players. *Proceedings of the National Academy of Sciences*, *113*(44), E6730–E6730. https://doi.org/10.1073/pnas.1614598113

Sullins, J. P. (2006). When is a Robot a Moral Agent. *International Review of Information Ethics*, *6*(12), 23–30.

Taylor, C., & Jefferson, D. (1993). Artificial Life as a Tool for Biological Inquiry. *Artificial Life*, *1*(1_2), 1–13. https://doi.org/10.1162/artl.1993.1.1_2.1

Tierney, H., & Glick, D. (2020). Desperately seeking sourcehood. *Philosophical Studies*, *177*(4), 953–970. https://doi.org/10.1007/s11098-018-1215-3

Tigard, D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 435–447. https://doi.org/10.1017/S0963180120000985

Tollon, F. (2021). Do Others Mind? Moral Agents Without Mental States. *South African Journal of Philosophy*, *40*(2), 182–194.

Tomasello, M. (2000). Two hypotheses about primate cognition. In *The evolution of cognition* (pp. 165–183). The MIT Press.

Tomasello, M. (2016). *A Natural History of Human Morality*. Harvard University Press.

Torrance, S. (2008). Ethics and Consciousness in Artificial Agents. *AI and Society*, *22*(4), 495–521. https://doi.org/10.1007/s00146-007-0091-8

Torrance, S. (2014). Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology*, *27*(1), 9–29. https://doi.org/10.1007/s13347-013-0136-5

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, *35*(1), 147–163. https://doi.org/10.1007/s00146-018-0845-5

Tye, M. (2011). *Consciousness Revisited: Materialism without Phenomenal Concepts*. The MIT Press.

Tyran, K. L., & Gibson, C. B. (2008). Is what you see, what you get? The relationship among surface- and deep-level heterogeneity characteristics, group efficacy, and team reputation. *Group & Organization Management*, *33*, 46–76. https://doi.org/10.1177/1059601106287111

Usher, M. (2020). Agency, Teleological Control and Robust Causation. *Philosophy and Phenomenological Research*, *100*(2), 302–324. https://doi.org/10.1111/phpr.12537

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001

van Dijk, H., van Engen, M. L., & van Knippenberg, D. (2012). Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organizational Behavior and Human Decision Processes*, *119*, 38–53. https://doi.org/10.1016/j.obhdp.2012.06.003

van Knippenberg, D., & Schippers, M. C. (2007). Work Group Diversity. *Annual Review of Psychology*, *58*(1), 515–541. https://doi.org/10.1146/annurev.psych.58.110405.085546

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & SOCIETY*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Veluwenkamp, H. (2022). Reasons for Meaningful Human Control. *Ethics and Information Technology*, *24*(4), 51. https://doi.org/10.1007/s10676-022-09673-8

Vermaas, P. E., Carrara, M., Borgo, S., & Garbacz, P. (2013). The design stance and its artefacts. *Synthese*, *190*(6), 1131–1152.

Vermaas, P. E., & Houkes, W. (2006). Technical functions: A drawbridge between the intentional and structural natures of technical artefacts. *Studies in History and Philosophy of Science Part A*, *37*(1), 5–18. https://doi.org/10.1016/j.shpsa.2005.12.002

Waal, F. B. M. D. (1997). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals* (Revised edition). Harvard University Press.

Waal, F. B. M. D. (2006). *Primates and philosophers: How morality evolved* (pp. xix, 209). Princeton University Press.

Waal, F. B. M. D. (2009). Veneer Theory. In *Veneer Theory* (pp. 3–58). Princeton University Press. https://doi.org/10.1515/9781400830336-003

Waal, F. B. M. D., Churchland, P. S., Pievani, T., & Parmigiani, S. (2014). *Evolved Morality: The Biology and Philosophy of Human Conscience*. https://philpapers.org/rec/DEWEMT

Walker, M. (2006). *A moral paradox in the creation of artificial intelligence: Mary Poppins 3000s of the world unite!* 23–28.

Wallace, R. J. (1998). *Responsibility and the Moral Sentiments:* Harvard University Press.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right From Wrong*. Oxford University Press.

Wallach, W., Allen, C., & Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness*, *03*(01), 177–192. https://doi.org/10.1142/S1793843011000674

Waller, R. R. (2014). The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument. *Philosophia*, *42*(1), 209–222. https://doi.org/10.1007/s11406-013-9476-8

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, *24*(2), 227–248.

Watson, G. (1999). Soft Libertarianism and Hard Compatibilism. *The Journal of Ethics*, *3*(4), 353–368. https://doi.org/10.1023/a:1009819618482

West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary Explanations for Cooperation. *Current Biology*, *17*(16), R661–R672. https://doi.org/10.1016/j.cub.2007.06.004

Wiegel, V. (2006). *Building blocks for artificial moral agents*.

Williams, K. Y., & O'Reilly, C. A. (1998). *Demography and diversity in organizations: A review of 40 years of research*.

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Harvard University Press.

Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. In F. D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46–62). Cambridge University Press.

Wolf, S. (1993). *Freedom within Reason*. Oxford University Press.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347. https://doi.org/10.1111/phpr.12095

Wylie, A. (2012). Feminist Philosophy of Science: Standpoint Matters. *Proceedings and Addresses of the American Philosophical Association*, *86*(2), 47–76.

Wylie, A. (2015). A Plurality of Pluralisms: Collaborative Practice in Archaeology. In F. Padovani, A. Richardson, & J. Y. Tsou (Eds.), *Objectivity in Science: New Perspectives from Science and Technology Studies* (pp. 189–210). Springer International Publishing. https://doi.org/10.1007/978-3-319-14349-1_10