Luise Müller*

Domesticating Artificial Intelligence

https://doi.org/10.1515/mopp-2020-0054
Published online February 14, 2022

Abstract: For their deployment in human societies to be safe, AI agents need to be aligned with value-laden cooperative human life. One way of solving this "problem of value alignment" is to build moral machines. I argue that the goal of building moral machines aims at the wrong kind of ideal, and that instead, we need an approach to value alignment that takes seriously the categorically different cognitive and moral capabilities between human and AI agents, a condition I call *deep agential diversity*. Domestication is the answer to a similarly structured problem: namely, how to integrate nonhuman animals that lack moral agency safely into human society and align their behavior with human values. Just like nonhuman animals, AI agents lack a genuinely moral agency; and just like nonhuman animals, we might find ways to train them to nevertheless assist us, and live and work among us – to "domesticate" them, in other words. I claim that the domestication approach does well in explaining many of our intuitions and worries about deploying AI agents in our social practices.

Keywords: AI ethics, philosophy of AI, value alignment, political philosophy, animal ethics

1 Introduction

In many of our social practices, AI agents are being developed and deployed: some examples are healthcare, transportation, armed conflict, and even love and friendship. Consider the possibility of robotic nurses or autonomous weapons; or consider recent journalistic and scholarly investigations of discriminatory algorithmic decision-making in policing, sentencing, and medical diagnostics (for examples, see Angwin et al. 2016; Thomas 2021). In all of these examples, there is a real possibility that the decisions of AI agents can undermine the values we usually take to be important for these practices: robotic nurses failing to respect the autonomy and bodily integrity of patients while caring for them, autonomous

***Corresponding author: Luise Müller**, Faculty of Humanities, Institute of Philosophy, University of Hamburg, Überseering 35, Postbox #4, 22297 Hamburg, Germany, E-mail: luise.mueller@uni-hamburg.de. https://orcid.org/0000-0001-8949-4825

weapons harming human bystanders and civilians, and algorithms used for policing, sentencing, hiring or medical diagnostics making decisions with discriminatory implications for protected classes. Thus, there is an increasing sense that AI agents – machines powered by artificial intelligence – must be equipped with special capabilities to make their deployment in human societies safe. Deploying AI agents safely means that we need to make sure that their decision making is not only aligned with our goals and preferences: as AI agents become more autonomous and operate at increasing speeds (Gabriel 2020, p. 412), humans need to ensure that the actions of AI agents are aligned to human values.

Thus, we have good reason to make value alignment of AI agents an important goal for their engineering, design, and regulation. But with that aim, at least two problems appear: first, *what values* do AI agents need to be aligned with? Any decision about which moral principles or which moral values AI agents ought to apply is more difficult than may appear at first, because in the real world there is disagreement about which values are correct (see also Himmelreich 2020 and Misselhorn 2018). Reasonable people disagree about what theory best describes which values play what role in our social practices, and what theory should guide our behavior in moral contexts. Such a "reasonable disagreement" or "reasonable pluralism" is a defining feature of liberal democratic societies (Rawls 1993).

The problem of disagreement also continues beyond the theoretical level: not only do reasonable people disagree about correct values and theories of morality, they also disagree in their evaluations of practical moral problems, as controversies around issues like abortion or vegetarianism show. There are many practical moral issues on which there are sound arguments on both sides of the debate. And this is not only true for decision-making within moral contexts, but also for the *identification* of moral contexts, problems, and values in the first place: deciding whether some situation constitutes a moral problem is itself a value statement, and so is defining the criteria for identifying moral contexts and problems.

The second issue arising for the aim of value alignment is: *How* do we align AI agents' decisionmaking with human values? To appreciate the complexity of this issue, notice that the value alignment problem is not only about aligning the machines' behavior with the values of its user: AI agents will usually act on behalf of persons, but will thereby often impact other persons. So the problem is not merely that AI agents are unaligned to the goals and values of the person or company that deploy it, but that AI agents are unaligned to the goals and values prevalent in the social practices they are deployed in. There is a social dimension to the problem of value alignment that applies to many of the areas in which AI agents are (prospectively) deployed, because human practices are constituted by a web of interdependent decisions and actions, in which AI agents increasingly play a role. While AI research "has focused on improving the individual

intelligence of agents and algorithms" (Dafoe et al. 2020, p. 3), solving the value alignment problem requires engaging AI researchers to improve the *social* intelligence of AI agents on the one hand, and constructing an adequate theoretical picture of the value-laden relations humans and nonhuman agents engage in, on the other.

The latter task – theorizing normative relations – is one of the core competences of moral and political philosophers. In this paper, I argue that in order to develop an adequate picture of the normative relations between humans and AI agents, a novel perspective is warranted that lets us see the problem of value alignment in a new light. Where humans and AI agents interact and cooperate in social practices, the problem of value alignment arises under the conditions of what I call *deep agential diversity*: human-nonhuman constellations that encompass radically heterogeneous agents with categorically different cognitive and moral capabilities.

I begin the argument by describing the concept of deep agential diversity, and how it differs from various forms of human diversity (Section 2). I then discuss two approaches to the problem of value alignment under deep agential diversity. The approach of building artificial "moral" agents seeks to emulate human moral agency in AI agents, building them to resemble human agents in regard to their moral capabilities (Section 3). This approach however fails, because human morality is more demanding than what we can expect to represent in AI agents. Instead, I defend the alternative approach of "domestication", and argue that we can learn about value alignment for AI agents by learning from how humans managed to "value-align" another type of nonhuman intelligent agent, namely nonhuman animals (Section 4). I then sketch what principles and categories this might include, and consider some implications of my argument (Section 5). Finally, I conclude the article with a brief summary (Section 6).

2 Value Alignment Under Deep Agential Diversity

The relations of human agents are characterized by the equal moral standing of persons (what Kymlicka (2002) calls "the egalitarian plateau of modern political philosophy"). Moral equality (or equality in status) is the most fundamental principle in modern political and moral philosophy – it is perhaps *the* defining characteristic of human relations that they are *relations between moral equals*. This in turn restricts the set of options of how we may permissibly treat one another, and how we may distribute the benefits and burdens of social cooperation. It is morally unacceptable to have a society operating according to a distinction between masters and slaves, because it disregards our fundamental moral equality.

Another premise in moral and political philosophy is that humans are sentient and vulnerable beings that feel pain, pleasure, happiness and grief, and care about how their lives go. Humans normally have an encompassing interest in their own wellbeing. That is why we reject and disapprove when humans are treated badly or when their interests in wellbeing are set back unjustly. Our core normative concepts like rights, harm, wellbeing, oppression, and so on only make sense under the condition that humans are vulnerable and care about their wellbeing. And connected to this is a third premise worth mentioning in this context: namely, the fact that humans are, for the most part, moral beings, that is, they are capable of moral reasoning and behavior. Their moral capabilities include their being able to act according to moral reasons, to be bearers of moral duties, and to be morally responsible for their behavior.[1]

These three premises – that humans are moral equals, that they are vulnerable, and that they possess moral agency – are among the basic preconditions for our theories about what is morally right, ethical, or just between humans. Any theory that fails to affirm or recognize these premises would get the normative dimension of human relations wrong: racist, sexist, or caste-based theories reject fundamental human moral equality, and are therefore unacceptable. Some political visions view persons as mere pawns or means to an abstract political goal, and are blind to the wellbeing of persons. And theories that ignore our moral capacities treat persons as if they were children at best, and irrational creatures to be controlled and manipulated at worst. None of these theories are acceptable by our normative standards, because they are fundamentally misleading about who we are as humans, and how we relate to one another. When we evaluate and regulate our social relations, these foundational premises are crucial.

But the fact that theories about social relations are typically developed with humans in mind is a problem for theorizing the ethical deployment of AI agents in human societies.[2] When we think of our moral or political relations, we think of agents who look, walk, and talk like you and me. In that sense, we are "unreflective speciesists". And because of that, we lack the methodological tools to understand social systems that are characterized by what I want to call deep agential diversity. The term denotes the property of social systems that contain human as well as nonhuman agents. Within social systems characterized by deep agential diversity, agents cooperate and work together in a number of different constellations: first,

1 Carissa Veliz (2021) even argues that the fact that humans are sentient is constitutive of the human moral capacity: "I claim we will not get an agent that can identify ethical problems and respond adequately to them without sentience" (p. 3).

2 There are some notable examples which are mostly in the postmodernist/poststructuralist tradition, see the work of Haraway (2016) or Latour (1993).

and obviously, humans cooperate with other humans; second, human agents also now increasingly cooperate with AI agents; and third, AI agents also cooperate with one another. This results in a complex web of interrelated actions that are increasingly transforming human social practices as we know them.

Such a constellation is diverse, because agents do not simply differ in their values and goals – which is a feature of social cooperation more generally – but they also differ categorically in their agential capabilities, vulnerabilities, and moral standing. While humans have equal moral standing, AI agents have no moral standing at all – in the sense that they matter morally for their own sake. Also, AI agents are not vulnerable in the same sense as humans are vulnerable – they have no wellbeing that can be threatened. And finally, AI agents possess a different set of agential capabilities than humans. In this regard, it is particularly noteworthy that AI agents lack moral agency: they are not thought to be able to bear moral duties and responsibility. In addition to these categorical differences between humans and AI agents, there are significant differences in complexity and agential quality among AI agents. What AI agents are optimized for depends on their respective context of deployment: whether they are autonomous weapon systems, care robots, language models, or Go-playing algorithms. This means that AI agents are only capable of operating in a limited domain, and mostly within the context of the practice they are deployed in.

This is why such constellations are a puzzle for moral and political philosophers: we lack adequate concepts that enable us to analytically capture a social system that is characterized by deep agential diversity. The methods currently employed by political philosophers are inapt for developing normative concepts of moral, fair, or just cooperation that prove applicable to societies in which humans and machines interact and cooperate in a meaningful manner. The problem of deep agential diversity thus induces us to explore and develop normative theories that are adequate for social systems that are "populated" by different kinds of agents exhibiting heterogeneity in abilities, autonomy, moral capability, moral status, and vulnerability.

One might object that human relations themselves are characterized by what I called deep agential diversity: while we accept that all humans have equal moral standing, humans do differ in their moral capabilities and their vulnerabilities. Hence, introducing AI agents into our system of social cooperation should not be a new challenge. What warrants my claim that deep agential diversity is indeed a problem and calls for a new form of theorizing?

It is indeed true that human relations are characterized by a variety of differences: humans have different cultures, politics, and goals; they have different social and class status; they play different roles in social life; they have different talents and intelligences, some have more power and knowledge, and so on.

Recognizing these differences and theorizing their impact on our social order is, as mentioned above, the core task of political and moral philosophy. And it is also true that humans differ in their vulnerability and moral capabilities: children, for example, are typically more vulnerable than adults, because their physical and cognitive abilities are still developing, and they depend on adults to foster these capacities. At the same time, children's moral capabilities are characteristically not fully developed until they reach a certain age. That is why we hold their parents to be responsible not only for ensuring their wellbeing, but also for teaching them the fundamental moral rules and principles that govern societies. In that sense, human parent–child relations exhibit agential diversity because the agential capabilities regarding moral agency differ between them: Parents are held to be full moral agents, whereas children are not yet full moral agents.

But this diversity is only on the surface level, because children neither lack moral capabilities, nor do they exhibit a categorically different kind of moral capability than adult humans. Rather, their capabilities are in development. Moral agency, one could say, is a "range property" (Rawls 1999, p. 444): it allows for variation within a certain range. The same is true for human vulnerability: children, the elderly, the chronically ill, and persons with serious cognitive impairments are certainly *more* vulnerable than adult humans with fully able bodies and minds. But again, although not all persons may have exactly equal vulnerability, they are equally vulnerable in some way or another. Vulnerability is also a range property that persons exhibit to various degrees.

It is in these three regards – moral capability, vulnerability, and moral status – that the agential diversity of the relations between humans and AI agents is "deep". While humans among themselves differ significantly from each other in the first two respects, they only differ in range, not in kind. What makes the diversity between humans and AI agents deep is the categorical nature of its difference.

3 The "Moral Machine" Approach

One possible strategy to tackle this problem might be to reduce the deep agential diversity in social systems by attempting to build AI agents that resemble humans in their cognitive and behavioral aspects. In particular, there is the idea that AI agents can be moral agents (Floridi and Sanders 2004), and that such AI agents capable of making moral decisions could be built (Anderson and Anderson 2007; Misselhorn 2018; Wallach and Allen 2009), or serve as a regulatory ideal (Wallach and Vallor 2020, p. 394). Where AI agents are used and deployed in contexts in which more or less autonomous moral decisionmaking is required, those AI agents

should be built in a way that makes their decisionmaking safe for humans; that is, those AI agents should be designed so as to make correct moral decisions. This can be done in a number of ways, and commonly cited distinctions in the literature are between top–down, bottom–up, hybrid, and virtue ethics approaches to building moral machines.

Some have called into question whether AI agents can in principle be real moral agents: without sentience, Veliz (2021) argues, algorithms are moral zombies rather than genuine moral agents; or because moral decisions are deeply personal and require moral emotions such as remorse, Sparrow (2020) argues that machines cannot be moral. I am sympathetic to these arguments, and I agree with these authors that we cannot expect machines to be moral agents in the same "full" sense that we think humans are moral agents. Even though the use of the term "moral" evokes association with human morality, AI agents lack access to many of the concepts we usually think of as components of morality, like empathy, a sense of justice, feelings of remorse or guilt, or an oppositional normative pull in the face of moral dilemmas.

But it is not clear that these arguments actually preclude what those arguing for moral machines suggest. Notice that the proponents do not claim that machines can be made fully moral in the human sense. Instead, they operate with a watered-down conception of moral agency. For example, consider how Anderson and Anderson begin one of their articles:

> The ultimate goal of machine ethics, we believe, is to create a machine that *itself* follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take. (2007, p. 15)

Or when Wallach and Allen assert that

> moral agents monitor and regulate their behavior in light of the harms their actions may cause or the duties they may neglect. Humans should expect nothing less of AMAs. A good moral agent is one that can detect the possibility of harm or neglect of duty, and can take steps to avoid or minimize such undesirable outcomes. … Perhaps even the most sophisticated AMAs will never really be moral agents in the same sense that human beings are moral agents. But wherever one comes down on the question of whether a machine can be genuinely ethical … an engineering challenge remains: how to get AI agents to act as if they are moral agents. (2009, pp. 26f.)

Neither Anderson and Anderson, nor Wallach and Allen claim that machines are morally capable of doing what humans can do. Perhaps there is simply no reason to claim that agents must be moral in a full and demanding human-like sense in order to be considered artificial "moral" agents. Artificial moral agents might not be sentient, or be able to have moral emotions like remorse, but they *can* (or can be

made to) act upon moral principles and rules, which is what matters in practical social contexts. AI agents, as Wallach and Allen point out in the above quote, must merely act *as if* they are moral agents. And that seems correct: when it comes to social life, the relevant ability is being able to act according to moral principles and moral rules.

Hence, we might not actually need a principled argument about why we should under no circumstances call AI agents "moral" agents – except perhaps to avoid any implication that artificial moral agents (AMA) carry moral and, by extension, legal responsibility. Indeed, this seems to be the main reason why many scholars warn or argue against "artificial moral agents" (Bryson, Diamantis, and Grant 2017, p. 282; Bryson 2018; Véliz 2021, p. 3). However, conceptually, moral agency and moral responsibility can be held apart. That is, we can conceptually conceive of AI agents that are capable of acting according to moral rules and principles, but that carry no moral responsibility. Indeed, this seems to be an adequate and more realistic picture of the role of artificial moral agents. And it is in fact how we conceive of human children: we know that they are capable of acting according to moral rules and principles (and we stimulate and encourage this capacity when we raise them), but we also know that until they reach a certain point of maturity, they cannot be held morally responsible for most of their behavior. Thus, one might think that artificial moral agents are best conceived of in analogy to human children (Misselhorn 2018, p. 165).

But that is still an inapt analogy, and I argue that we understand the differences between human moral agents and AI agents better when we conceive of AI agents in analogy to nonhuman animals. One of the core insights of the literature in the philosophy and theory of animal rights is that nonhuman and human animals are distinct in their moral capabilities: while many animals have considerable social skills – which is one of the reasons why humans choose to live or work with them – they lack full moral agency. Consider dogs: dogs are not only popular companion animals, they also work with humans in a number of contexts. They look after sheep, they are used as guide dogs, they do police work, and so on. Dogs can be taught to live safely among humans in their homes, they can be trained to exhibit behavior that is aligned to human needs and habits, and they exhibit the skills to navigate human contexts so reliably and competently that we have become used to seeing dogs in many domains and areas of social life. In human contexts, dogs are successfully trained to submit to human principles and rules without themselves understanding, reflecting, or affirming the point of these practices. They are the addressees of normative rules, without being their authors.

Something similar is true for the distinction between human moral capacities and artificial moral capacities: humans are not only capable of acting according to moral rules and principles, but over and above, they are also capable of

constructively deliberating and questioning those rules, by asking whether they are an adequate normative reflection of any human's moral standing within their social relations. Human moral values are not a fixed system, but subject to challenge, reflection, evaluation, revision, and reform. Human moral agents do not simply submit to moral, social, and political practices blindly, but reflect on them and potentially induce change. In other words, they are the common authors of the rules they live under, and they reciprocally recognize each other as such.

Just like nonhuman animals, AI agents are incapable of taking such a reflexive standpoint,[3] or what Christine Korsgaard (2010) calls "normative self-governance": which is the "capacity to assess the potential grounds of our beliefs and actions, to ask whether they constitute good reasons, and to regulate our beliefs and actions accordingly" (p. 6). To be a normatively self-governing creature, agents must be able to recognize and act upon moral principles for the right reasons instead of mechanically applying them.

That is why teaching moral decisionmaking in machines is disanalogous to teaching one's child to be moral. The analogy is misleading because human children are raised to be full moral agents: they are expected to not only apply their training, but to take a reflexive standpoint and to act for the right reasons. To mark this distinction between human moral agency and artificial agency, let us call the morally relevant capacities that AI agents can realistically acquire *moral competence*. Morally competent agents can regulate and alter their behavior on the basis of its normatively significant impact on others. Their actions, decisions, or behavior are motivated by training, habit, coercion, or manipulation.

Why does the distinction between artificial moral competence and human moral agency matter? Recall that the problem we are interested in is not whether we may or may not attribute the term "moral agents" to AI agents, but whether the idea of building moral machines can solve the problem of value alignment under deep agential diversity. My argument is that building moral machines with the goal of emulating human moral agency focusses on the wrong idea of morality. Recall that I emphasized that AI agents are unable to normatively self-govern and take a reflexive standpoint towards moral rules and principles. AI agents are unable to act morally in a way that humans can: namely, by taking a reflexive standpoint and by being normatively self-governing. But these capacities are essential for the *autonomous* alignment of values: we humans deliberate about what we owe one

3 This is not to say that this is the only morality-related difference between human agents and AI agents, but my claim is that this capability is important for the problem of value alignment. For a discussion of a number of differences, see Misselhorn (2018, p. 164). Misselhorn also argues that the difference in moral agency between humans and artificial systems is a feature, rather than a bug, because artificial systems that use higher order reasoning to question moral rules would put users at risk.

another, how we ought to treat each other, and whether our social practices are in line with our moral and ethical values, or whether they require reform and change.

In other words, we engage in constant reciprocal value alignment by reflecting on our moral decisions and how they impact and relate to other persons. We ask whether the moral rules and principles we hold are normatively valid. As a kind of creature that can reflect upon the fact that they are capable of acting according to moral rules and principles, we are normatively self-governing in a way that AI agents cannot be: we are the authors of the normative rules and principles that govern our social practices, because we are not merely capable of acting according to moral rules and principles, but of taking a reflexive standpoint.

We can now see why the proposal to build moral machines fails as a strategy to reduce deep agential diversity. It falls into a dilemma: either it fails because it focusses on the wrong kind of ideal for AI agents – an ideal that attempts to emulate human moral agency, but one which AI agents will always fall short of. Human moral agency is distinguished by the capability to normatively self-govern, while AI agents are merely capable of what I called moral competence: of acting according to moral rules and principles by having been trained or manipulated to do so. This is the first horn of the dilemma.

The alternative is to operate with a watered-down conception of moral agency that is reachable for AI agents (and what I called moral competence). But this wrongly assumes that moral agency merely consists of being able to apply rules to a social context, or act according to moral rules. But morality is more demanding: it consists of reflecting, evaluating, deliberating, and collectively changing and reforming those moral rules and principles. Mere moral competence is insufficient for actually reducing deep agential diversity and all the issues that come with it, because only normatively self-governing agents can deliberate and evaluate moral rules and principles, and are therefore capable of aligning their values autonomously with others. This is the second horn of the dilemma.

In conclusion, the strategy to reduce deep agential diversity by building artificial moral agents – with the goal that the cognition and behavior of these AI agents emulates human moral behavior – either fails because it ascribes an unrealistic ideal of moral agency to AI agents, or it fails because it operates with a concept of moral agency that wrongly reduces moral agency to the application of moral rules and principles.

4 The Domestication Approach

Having seen that the "moral machine" approach to tackling deep agential diversity in social systems fails, I want to propose an alternative. The alternative strategy

approaches deep agential diversity by accepting it as a social fact, and by building that fact into our theory of how to align AI to human values. This means that we take seriously the limits of the moral capabilities of AI agents, and integrate this insight into a larger picture that adequately represents the diversity of the cognitive and moral capabilities of agents cooperating in our social practices.

I have already argued that we understand the limits of AI agents' moral capabilities better if we compare them with nonhuman animals. I now want to argue that in order to develop a useful and normatively accurate picture of our relations with nonhuman intelligent agents, we can also learn from our experience with nonhuman animals. This is because aligning AI agents to human values is structurally analogous to domesticating nonhuman animals: domestication allows human moral agents to cooperate with nonhuman agents without human-like moral capabilities. The agential qualities and cognitive capacities of animals differ radically from humans, and yet a very fruitful discussion about the normative relations between humans and animals has delivered insights about what roles nonhuman animals can play in human social systems, and what morally follows from those roles.[4]

Consider that the presence of nonhuman animals within human societies is the original value alignment problem: how do we integrate nonhuman animals into our human value-laden social practices in a way that is safe and efficient? The way humans have solved this problem is through domestication.[5] Zeder (2012) defines domestication as "a sustained, multigenerational, mutualistic relationship in which humans assume some significant level of control over the reproduction and care of a plant/animal in order to secure a more predictable supply of a resource of interest" (pp. 163f.). Animals are deliberately bred and selected for specific traits and behavioural dispositions that makes human-animal cooperation and co-living safe.

The goal of the domestication process was not to make animals into moral agents: we now look at the infamous animal trials[6] in the Middle Ages with a mix of amusement and incredulity, in which animals were treated as if they were moral agents responsible for their behavior. It now seems absurd to us to put horses, pigs,

4 See Anderson (2005); Donaldson and Kymlicka (2011); Garner (2013); Korsgaard (2018); Nussbaum (2006); Valentini (2014).

5 Of course, humans have also used other strategies – also violent strategies – to separate and keep themselves safe from animals, like culling, extermination, or territorial separation. Some animals, like many wild animals, are not fit for living among humans; their permanent presence is simply too dangerous. In Elizabeth Anderson's (2005) words, there are some nonhuman animals with whom peace is not possible (p. 288). We ban those animals from our dwellings and houses, and maybe this is indeed what we should do in the cases of some technologies that are simply too dangerous – ban them or "kill" them before they even get built. But my focus in this paper is on domestication, so I won't pursue this further.

6 See the work of Carson (1917).

or vermin on trial and sentence them to punishment for having killed a person, because animals bear no moral agency, and have no moral responsibility for their behavior. Unlike most adult humans, we do not expect animals to be able to make decisions for the right moral reasons – and we do not punish them if they fail to do so.

Consider what I have earlier called moral capability. As far as we know today, animals do not possess the full set of moral agential qualities (Tomasello 2009). We do not hold cats responsible for maiming and killing birds, because they lack the reflective capacities to grasp moral concepts like a duty not to kill. Nonhuman animals lack the capacity to reflect on and understand normative rules, and therefore we do not attribute responsibilities and duties to them.[7]

And yet, many nonhuman animals live among, and with, humans. Humans have "value-aligned" animals by having recognized the deep agential diversity between humans and animals as a fact, and having structured their normative relations to animals accordingly. The reason we keep certain kinds of companion animals, and not others, is that some have been selectively bred for the traits that allow them to live with humans. They live in our homes because they can be trained to respect human rules (like the rules that peeing on the couch is unacceptable, that bringing dead mice to human dinner is not welcomed, or that biting the postman's leg results in punishment and should be avoided). These animals are capable of learning and complying with rules, but they clearly do not do so out of a genuine moral motivation (they do not "respect" those rules because they have a rational insight that biting the postman's leg would be unethical, but because they have been trained or manipulated into obeying that rule).

Something similar applies to AI agents: AI agents do not understand moral rules and principles in the way humans do, but like animals, they probably can be "made" to be morally competent in the sense described above. They can regulate their behavior so that it aligns to human rules and principles without being capable of understanding the nature or purpose of these rules and principles. Moral competence in the sense described above is sufficient to navigate cooperative systems – as the analogy with animals shows, human-like moral agency is not necessary for cooperating with humans in a safe manner.

All of this is not to say that AI agents are *just like* nonhuman animals; they are different in many important ways. For one, the animal rights literature has shown that animals can be bearers of rights because many possess the capacity to subjectively experience their life and it therefore matters how they fare in it. As it stands, AI agents lack the moral standing needed to be a bearer of rights, because

7 Korsgaard (2018, pp. 58ff.) argues that we cannot apply moral standards to animal behavior, because those standards are external in a sense, and thus simply do not apply to them.

they neither possess the ability to feel pleasure and pain, nor do they possess subjective interests that would make their "existence" go better or worse: there is no one "home".[8]

Even though it is controversial in the animal ethics literature to what extent we can say that humans and animals are moral equals,[9] almost all ethicists and philosophers agree that we owe animals at least some moral consideration. This is of course different when it comes to AI agents. Even though new-generation conversational machine learning algorithms like GPT-3 may in some instances appear to have a subjective self or a personality, there is no indication whatsoever that it has any subjective interests that would need protection in the form of rights. Although the language skills of GPT-3 are impressive to users, Bender and Gebru (2021) argue that "no actual language understanding is taking place" in these large language models (p. 615). Why does that matter for value alignment? The fact that AI agents lack any moral status implies that AI agents do not matter morally for their own sake. As a consequence, and in contrast to our relations with animals, when we decide how we distribute the burdens and benefits of human-AI cooperation, we humans are morally permitted to offload the burdens to AI agents, and take all the benefits for ourselves.

5 Implications

My claim is that in contrast to the moral machine approach, the domestication approach to value alignment provides a more realistic picture of human-AI cooperation. It better fits many of our intuitions, and it better fits the worries about AI agents that we find in the current AI ethics literature. In this section, I develop a rough outline of the possible categories and principles that the domestication approach implies.[10] The idea is that the normative principles and values governing a social practice depend (a) on the type of agents involved, (b) on the kind of relationship human agents have with it, and (c) on the context of this relationship.

8 For an argument on how this could change in the long term, and that we might need at least some type of moral status for AI, see Risse (2019, pp. 22ff.).

9 See Cochrane (2018) and Ladwig (2020) for an argument affirming the moral equality between humans and animals, and Kagan (2019) for an argument against it.

10 Let me stress that this is merely a sketch, and that I do not intend this sketch to exhaust the possible and even plausible categories and principles for thinking about value alignment.

5.1 Type-Dependence

One insight gained from the animal analogy is that just as it makes little normative sense to treat animals as an undifferentiated mass, it makes little normative sense to treat all AI agents as an undifferentiated mass. The obligations we owe to animals are not all the same, but rather depend on the kind of animal. For example, the extent to which animals are owed rights depends on the nature of the animal: we are less inclined to ascribe rights to insects or shrimps than to, say, great apes, dolphins, or elephants. The reasoning behind this rationale is that the less cognitively complex an animal is, the less complex, deep, and rich is its subjective experience.

The principles needed to govern AI agents in human cooperative contexts will depend on the kind of AI agent involved and, in particular, on the degree of its autonomy. The more autonomous AI agents act or decide, the more cautious should our rules and principles governing the AI agent be.

5.2 Relation-Dependence

Normative principles are dependent on the kind of relationship agents have. Consider how we traditionally distinguish our normative rules and principles by the type of relation they apply to: for example, what we owe to our fellow countrymen and women is distinct from what we owe to persons living in other states. Of course, the extent to which these duties differ is subject to debate: some argue that we have very strong duties towards our fellow citizens, and very loose duties towards citizens from other countries; and others argue that our duties towards our fellow citizens and citizens of other states do not differ very much. But what is important is that we differentiate between various political and social relations. And by doing that, we also discern what rights and obligations appropriately apply to the roles agents inhabit in these relations.

This is also true for animal ethics: a new wave of literature very helpfully distinguishes between the different types of relations we humans have with animals. The relation in which we stand to our companion animals requires different rights (for the animals) and duties (for their "owners") than our relation to wild animals. Wild animals are, above all, owed that humans leave their territory intact, and do not interfere with their ecosystem. Donaldson and Kymlicka (2011) even use the analogy of wild animals as "sovereign nations", which brings out the point nicely. On the other hand, companion animals live in a relationship of dependence with humans, and we therefore owe them adequate shelter and food. Other animals in work contexts, like police or guide dogs or farm animals, often spend much

of their lives "working" with and for humans, so we might think that they are owed medical assistance when they need it, or a right to retire when they are too old to do their jobs.

The idea of relation-dependence strikes me as equally relevant for AI agents. Relations between humans and AI agents can range from complete control by humans over AI agents to relationships in which humans are more or less dependent on the AI agent. For example, in contexts of care, the relation between AI agents and users might be a relation of dependence, as is the case with social robots designed to assist patients with Alzheimer or dementia. Such users are much more vulnerable and are therefore in a position that warrants extra safety from exploitation, for example. In other contexts, such as in commercial or in an industrial setting, humans might have much firmer control over the AI agent. Again, the more dependent/vulnerable human users are, the more cautious should our governance be – with some cases of relationships where using AI agents will be outright ethically impermissible.

5.3 Context-Dependence

The idea of context-dependence denotes that for the distribution of rights and obligations, the context in which we deploy AI agents is important. In particular, the context will determine what values we emphasize. For example, to what extent we value privacy over precision might depend on whether the deployment context is algorithmic decisionmaking in hiring or criminal justice, or a medical digital twin that closely monitors the progression of a serious disease. Similarly, different normative principles and rules will apply to hiring algorithms and autonomous weapon systems: we might emphasize the value of non-discrimination and fairness in the one case, and the value of non-harming in the other.

What does my argument imply? The aim of this paper was not to provide an ethical recommendation about the kinds of values AI agents must consider, or in what form we should package these values to make them applicable to AI agents. The aim of this paper is rather to provide a framework that helps us to make better sense of our relations to AI agents. So in what sense does the domestication approach change our perspective on the design and deployment of AI agents?

In terms of future research, my arguments imply that our priorities should be on collectively working out the ethical implications and pitfalls of deploying specific AI agents in specific domains and for specific tasks – something which the AI ethics community is already vitally involved in. Emphasizing type-, relation-, and context-dependence goes against the idea that we can just formulate a number of principles for beneficial, human-centered, benign, safe, or ethical AI

that apply across the board.[11] Ethical challenges arise in specific contexts and within specific relations, and what we want to know is what it *means* to implement these values in particular human practices. What does it mean for algorithmic decision-making to be fair in predictive policing? Or what does it mean for healthcare AI to be beneficial? To whom? And which duties, responsibilities, and liabilities does this give rise to exactly? The type-, context-, and relation-dependence of these challenges explains why AI ethicists are exploring the different ethical contexts in which we deploy AI agents, and interpret and construct normative principles and rules in light of these specific contexts.

When it comes to practical questions, one implication of my arguments is that wherever AI agents are deployed, humans remain the sole bearers of responsibility: that is, humans are paternalistically responsible (and liable) for nonhuman agents. My earlier analysis of how human moral capabilities are different from artificial moral competence explains our intuitions why, on the one hand, humans are the sole authorities when it comes to morally relevant rules and principles, and why, on the other, humans are also the sole bearers of responsibility for the decisions of AI agents, even when they become increasingly autonomous. With other (rational and reasonable) humans, we get into debates about the normative principles and rules that should govern our societies, because we know that (above a certain quality threshold) their arguments count just as much as our arguments. As I said above, humans are in a constant process of aligning their values to each other by challenging and changing the normative rules that guide our social practices.

But we certainly don't do the same with animals: we do not seek agreement about rules regarding leg-biting, but manipulate and train them into obeying our normative rules. We adopt a paternalistic stance in which *we* decide what animals are permitted to do, and which rules apply. And if they deviate from these rules, we do not hold animals responsible, but their owners. To a reasonable extent, humans are responsible (and often also liable) for damage done by their companion animals or their farm animals.

I believe that a paternalistic stance is precisely what is adequate for AI agents: where humans and AI agents cooperate, humans are the sole authorities on what the rules are, and what permissions AI agents can be given. The analogy with animals also explains our intuitions regarding responsibility: to ensure safety and avoid responsibility gaps, autonomous AI agents must be assigned human bearers of responsibility and liability for damages caused by AI agents. In some instances, this may ultimately lead to a situation in which some AI agents become impermissibly costly, and are banned or phased out – which is exactly how we regulate

11 For a similar argument, see Mittelstadt (2019).

the keeping of dangerous animals: where the costs are too high, and the owners are incapable of ensuring safety for themselves and others, those animals are barred from being kept in human society.

That humans are paternalistically responsible (and liable) for nonhuman agents also implies that full automation of potentially harmful AI agents is undesirable. We might generally disencourage the development and deployment of AI agents in domains and contexts where decisionmaking with squarely moral implications is required. This might mean that we do not only ban technologies that might be dangerous if used for sinister purposes, but that we also do not automate processes that require genuine moral decisionmaking.

6 Conclusion

In this article, I focused on the problem of value alignment, and argued that the technological advances prompted by the development of artificial intelligence warrant a new perspective on our social relations: one that takes into account the deep agential diversity of our social systems. Where human and AI agents cooperate, that cooperation is characterized by heterogeneous types of intelligence and moral capabilities. Responding to the value alignment problem under these conditions, I argued that the ideal of building moral machines that emulate human moral capabilities focuses on the wrong goal, because it is either based on an unrealistic picture of what AI agents are capable of, or it operates with an inadequate concept of morality.

Instead, I claimed that an analogy to our relations to nonhuman animals provides a more plausible picture of how humans could relate to AI agents: while nonhuman animals and AI agents are different in many ways, the fundamental approach with which we have "value-aligned" animals is transferable to AI agents, and can guide us in understanding how human moral agents can cooperate with nonhuman agents without human-like moral capabilities. I then outlined three categories that should help us work out the normative principles and rules governing a social practice: I argued that those principles and values are dependent on the kind of agent, the nature of the relationship to the agent, and the deployment context of that agent.

In political philosophy, we are generally interested in how it is possible to preserve the equality and freedom of persons in the face of a set of given natural and social circumstances. Part of the social circumstances are technological advances: they impact – and sometimes transform – human relations, create new opportunities for flourishing and independence, but also for exploitation and dependency. The approach I defended and the resulting framework I developed in

this article might give us some orientation on how to begin thinking about these challenges more rigorously.

Acknowledgments: For a very constructive first discussion of this paper, I would like to thank our informal philosophy of AI working group: Birgit Beck, Hauke Behrendt, Johannes Himmelreich, Janina Loh, Wulf Loh, and Julian Müller. I am also grateful to Peter Niesen and the participants of his political theory colloquium for discussing the paper with me at a later stage. Finally, I received immensely helpful comments from two anonymous reviewers, who have provided thoughtful and detailed comments, which helped me improve the paper enormously.

References

Anderson, E. 2005. "Animal Rights and the Values of Nonhuman Life." In *Animal Rights: Current Debates and New Directions*, edited by M. Nussbaum, and C. Sunstein, 277–98. Oxford: Oxford University Press.

Anderson, M., and S. Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28 (4): 15–26.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. "Machine Bias." *Pro Publica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed January 10, 2022).

Bender, E., and T. Gebru, et al. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–23, https://doi.org/10.1145/3442188.3445922.

Bryson, J. 2018. "Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20: 15–26.

Bryson, J. J., M. Diamantis, and T. D. Grant. 2017. "Of, for, and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25: 273–91.

Carson, H. L. 1917. "The Trial of Animals and Insects: A Little Known Chapter of Medieval Jurisprudence." *Proceedings of the American Philosophical Society* 56 (5): 410–5.

Cochrane, A. 2018. *Sentientist Politics: A Theory of Global Inter-Species Justice*. Oxford: Oxford University Press.

Dafoe, A., E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel. 2020. "Open Problems in Cooperative AI." arXiv:2012.08630v1.

Donaldson, S., and W. Kymlicka. 2011. *Zoopolis: A Political Theory of Animal Rights*. Oxford: Oxford University Press.

Floridi, L., and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds & Machines* 14: 349–79.

Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–37.

Garner, R. 2013. *A Theory of Justice for Animals*. Oxford: Oxford University Press.

Haraway, D. 2016. *Staying with the Trouble: Making Kin in the Cthulucene*. North Carolina: Duke University Press.

Himmelreich, J. 2020. "Ethics of Technology Needs More Political Philosophy." *Communications of the ACM* 63 (1): 33–5.
Korsgaard, C. 2010. "Reflections on the Evolution of Morality." *The Amherst Lecture in Philosophy* 5: 1–29.
Korsgaard, C. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.
Kymlicka, W. 2002. *Contemporary Political Philosophy: An Introduction*. Oxford: Oxford University Press.
Latour, B. 1993. *We Have Never Been Modern*. Massachusetts: Harvard University Press.
Ladwig, B. 2020. *Politische Philosophie der Tierrechte*. Berlin: Suhrkamp.
Mittelstadt, B. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1: 501–7.
Misselhorn, C. 2018. "Artificial Morality: Concepts, Issues, Challenges." *Society* 55: 161–9.
Nussbaum, M. 2006. *Frontiers of Justice: Disability, Nationality, Species Membership*. Massachusetts: Harvard University Press.
Rawls, J. 1993. *Political Liberalism*. New York: Columbia University Press.
Rawls, J. 1999. *A Theory of Justice, Revised Edition*. Massachusetts: Harvard University Press.
Risse, M. 2019. "Human Rights, Artificial Intelligence, and Heideggerian Technoskepticism: The Long (Worrisome?) View". Carr Center Discussion Paper (CCDP) 2019–002.
Sparrow, R. 2020. "Why Machines Cannot Be Moral." *AI & Society* 36 (3): 685–93.
Thomas, R. 2021. "Medicine's Machine Learning Problem." *Boston Review*. http://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem (accessedJanuary 10, 2022).
Tomasello, M. 2009. *Why We Cooperate*. Massachusetts: MIT Press.
Valentini, L. 2014. "Canine Justice: an Associative Account." *Political Studies* 62 (1): 37–52.
Véliz, C. 2021. "Moral Zombies: Why Algorithms are Not Moral Agents." *AI & Society* 36: 487–97.
Wallach, W., and C. Allen. 2009. *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
Wallach, W., and S. Vallor. 2020. "Moral Machines. From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, edited by S. M. Liao. Oxford: Oxford University Press.
Zeder, M. 2012. "The Domestication of Animals." *Journal of Anthropological Research* 68 (2): 161–90.