4

Risk Imposition by Artificial Agents

The Moral Proxy Problem

Johanna Thoma[*]

I. INTRODUCTION

It seems undeniable that the coming years will see an ever-increasing reliance on artificial agents that are, on the one hand, autonomous in the sense that they process information and make decisions without continuous human input, and, on the other hand, fall short of the kind of agency that would warrant ascribing moral responsibility to the artificial agent itself. What I have in mind here are artificial agents such as self-driving cars, artificial trading agents in financial markets, nursebots, or robot teachers.[1] As these examples illustrate, many such agents make morally significant decisions, including ones that involve risks of severe harm to humans. Where such artificial agents are employed, the ambition is that they can make decisions roughly as good as or better than those that a typical human agent would have made in the context of their employment. Still, the standard by which we judge their choices to be good or bad is still considered human judgement; we would like these artificial agents to serve human ends.[2]

Where artificial agents are not liable to be ascribed true moral agency and responsibility in their own right, we can understand them as acting as proxies for human agents, as making decisions on their behalf. What I will call the 'Moral Proxy Problem' arises because it is often not clear for whom a specific artificial agent is acting as a moral proxy. In particular, we need to decide whether artificial agents should be acting as proxies for what I will call low-level agents – for example individual users of the artificial agents, or the kinds of individual human agents artificial agents are usually replacing – or whether they should be moral proxies for what I will call high-level agents – for example designers, distributors, or regulators, that is, those who can

[*] I received very helpful feedback on an earlier draft of this paper from Kate Vredenburgh, Silvia Milano, and Johannes Himmelreich. Previous versions of this paper were also presented at the University of York, at the Global Priorities Institute at Oxford, at the Third Workshop on Decision Theory and the Future of Artificial Intelligence at the Australian National University, at the Humanities and Social Sciences Colloquium at the Freiburg Institute for Advanced Studies (FRIAS), and at the Interdisciplinary Research Symposium on Global Perspectives on Responsible AI organised by the FRIAS Saltus Group on Responsible AI. I have benefitted immensely from discussions both at these events, as well as during my time as a visiting FRIAS Fellow in April 2019.

[1] See M Wellman and U Rajan, 'Ethical Issues for Autonomous Trading Agents' (2017) 27 *Minds and Machines* 609; A Sharkey and N Sharkey, 'Granny and the Robots: Ethical Issues in Robot Care for the Elderly' (2012) 14 *Ethics and Information Technology* 27; and A Sharkey, 'Should We Welcome Robot Teachers?' (2016) 18 *Ethics and Information Technology* 283 respectively for critical discussion of these types of agents.

[2] Note that I don't mean to restrict human ends to human interests in a narrow sense here. Insofar as humans can, and often do, have ends that are not speciesist, we can think of artificial agents being deployed to further such ends, for example in wildlife preservation.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

potentially control the choice behaviour of many artificial agents at once. I am particularly interested in the Moral Proxy Problem insofar as it matters for decision structuring when making choices about the design of artificial agents. Who we think an artificial agent is a moral proxy for determines from which agential perspective the choice problems artificial agents will be faced with should be framed:[3] should we frame them like the individual choice scenarios previously faced by individual human agents? Or should we, rather, consider the expected aggregate effects of the many choices made by all the artificial agents of a particular type all at once?

Although there are some initial reasons (canvassed in Section 2) to think that the Moral Proxy Problem and its implications for decision structuring have little practical relevance for design choices, in this paper I will argue that in the context of risk the Moral Proxy Problem has special practical relevance. Just like most human decisions are made in the context of risk, so most decisions faced by artificial agents involve risk:[4] self-driving cars can't tell with complete certainty how objects in their vicinity will move, but rather make probabilistic projections; artificial trading agents trade in the context of uncertainty about market movements; and nursebots might, for instance, need to make risky decisions about whether a patient symptom warrants raising an alarm. I will focus on cases in which the artificial agent can assign precise probabilities to the different potential outcomes of its choices (but no outcome is predicted to occur with 100% certainty). The practical design choice I am primarily concerned with here is how artificial agents should be designed to choose in the context of risk thus understood, and in particular whether they should be programmed to be risk neutral or not. It is for this design choice that the Moral Proxy Problem turns out to be highly relevant.

I will proceed by, in Section III, making an observation about the standard approach to artificial agent design that I believe deserves more attention, namely that it implies, in the ideal case, the implementation of risk neutral pursuit of the goals the agent is programmed to pursue. But risk neutrality is not an uncontroversial requirement of instrumentally rational agency. Risk non-neutrality, and in particular risk aversion, is common in choices made by human agents, and in those cases is intuitively neither always irrational, nor immoral. If artificial agents are to be understood as moral proxies for low-level human agents, they should emulate considered human judgements about the kinds of choice situations low-level agents previously found themselves in and that are now faced by artificial agents. Given considered human judgement in such scenarios, often exhibits risk non-neutrality, and in particular risk aversion; artificial agents that are moral proxies for low-level human agents should do so too, or should at least have the capacity to be set to do so by their users.

Things look differently, however, when we think of artificial agents as moral proxies for high-level agents, as I argue in Section IV. If we frame decisions from the high-level agential perspective, the choices of an individual artificial agent should be considered as part of an aggregate of many similar choices. I will argue that once we adopt such a compound framing, the only reasonable approach to risk is that artificial agents should be risk neutral in individual choices, because this has almost certainly better outcomes in the aggregate. Thus, from the high-level agential perspective, the risk neutrality implied by the standard approach appears justified. And so, how we resolve the Moral Proxy Problem is of high practical importance in the context of risk. I will return to the difficulty of addressing the problem in Section V, and also argue there

[3] Here and throughout, I use 'framing' in a non-pejorative sense, as simply referring to the way in which a decision problem is formulated before it its addressed.

[4] Frequently neglecting the context of risk is indeed a serious limitation of many discussions on the ethics of AI. See also S Nyholm and J Smids, 'The Ethics of Accident-Algorithms for Self-Driving Cars; an Applied Trolley Problem?' (2016) 19 *Ethical Theory and Moral Practice* 1275 (hereafter Nyholm and Smids, 'Ethics of Accident-Algorithms').

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

that the practical relevance of agential framing is problematic for the common view that responsibility for the choices of artificial agents is often shared between high-level and low-level agents.

II. THE MORAL PROXY PROBLEM

Artificial agents are designed by humans to serve human ends and/or make decisions on their behalf, in areas where previously human agents would make decisions. They are, in the words of *Deborah Johnson* and *Keith Miller* 'tethered to humans'.[5] At least as long as artificial agents are not advanced enough to merit the ascription of moral responsibility in their own right, we can think of them as 'moral proxies' for human agents,[6] that is, as an extension of the agency of the humans on whose behalf they are acting. In any given context, the question then arises who they should be moral proxies for. I will refer to the problem of determining who, in any particular context, artificial agents ought to be moral proxies for as the 'Moral Proxy Problem'. This problem has been raised in different forms in a number of debates surrounding the design, ethics, politics, and legal treatment of artificial agents.

Take, for instance, the debate on the ethics of self-driving cars, where *Sven Nyholm* points out that before we apply various moral theories to questions of, for example, crash optimisation, we must settle on who the relevant moral agent is.[7] In the debate on value alignment – how to make sure the values advanced AI is pursuing are aligned with those of humans[8] – the Moral Proxy Problem arises as the question of whose values AI ought to be aligned with, especially in the context of reasonable disagreement between various stakeholders.[9] In computer science, *Vincent Conitzer* has recently raised the question of 'identity design', that is, the question of where one artificial agent ends and another begins.[10] He claims that how we should approach identity design depends at least in part on whether we want to be able to assign separate artificial agents to each user, so that they can represent their users separately, or are content with larger agents that can presumably only be understood as moral proxies for larger collectives of human agents. Finally, in debates around moral responsibility and legal liability for potential harms caused by artificial agents, the Moral Proxy Problem arises in the context of the question of which human agent(s) can be held responsible and accountable when artificial agents are not proper bearers of responsibility themselves.

For the purposes of my argument, I would like to distinguish between two types of answers to the Moral Proxy Problem: on the one hand, we could think of artificial agents as moral proxies for what I will call 'low-level agents', by which I mean the types of agents who would have faced the individual choice scenarios now faced by artificial agents in their absence, for example, the individual users of artificial agents such as owners of self-driving cars, or local authorities using artificial health decision systems. On the other hand, we could think of them as moral proxies for

[5] DG Johnson and KW Miller, 'Un-Making Artificial Moral Agents' (2008) 10 *Ethics and Information Technology* 123.

[6] See J Millar, 'Technology as Moral Proxy Autonomy and Paternalism by Design' (2015) 34 *IEEE Technology and Society Magazine* 47. Also see K Ludwig, 'Proxy Agency in Collective Action' (2014) 48 *Noûs* 75 for a recent analysis of proxy agency, J Himmelreich, 'Agency and Embodiment: Groups, Human–Machine Interactions, and Virtual Realities' (2018) 31 *Ratio* 197 on proxy agency as disembodied agency and S Köhler, 'Instrumental Robots' (2020) 26 *Science and Engineering Ethics* 3121 on artificial agents as 'instruments' for human agents.

[7] S Nyholm, 'The Ethics of Crashes with Self-Driving Cars: A Roadmap, I' (2018) 13(7) *Philosophy Compass* 6.

[8] See, e.g. S Russell, *Human Compatible: AI and the Problem of Control* (2019) for a prominent book-length treatment.

[9] See, e.g. I Gabriel, 'Artificial Intelligence, Values and Alignment' (2020) 30 *Minds and Machines* 411 for discussion.

[10] V Conitzer, 'Designing Preferences, Beliefs, and Identities for Artificial Intelligence' (2020) 33(1) *Proceedings of the AAI Conference on Artificial Intelligence (hereafter Conitzer, 'Designing Preferences').*

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

what I will call 'high-level agents', by which I mean those who are in a position to potentially control the choice behaviour of many artificial agents,[11] such as designers of artificial agents, or regulators representing society at large.

I would also like to distinguish between two broad and connected purposes for which an answer to the Moral Proxy Problem is important, namely, ascription of responsibility and accountability on the one hand, and decision structuring for the purposes of design choices on the other. To start with the first purpose, here we are interested in who can be held responsible, in a backward-looking sense, for harms caused by artificial agents, which might lead to residual obligations, for example, to compensate for losses, but also who, in a forward-looking sense, is responsible for oversight and control of artificial agents. It seems natural that in many contexts, at least a large part of both the backward-looking and forward-looking responsibility for the choices made by artificial agents falls on those human agents whose moral proxies they are.

My primary interest in this paper is not the question of responsibility ascription, however, but rather the question of decision structuring, that is, the question of how the decision problems faced by artificial agents should be framed for the purposes of making design choices. The question of who is the relevant agent is in a particular context is often neglected in decision theory and moral philosophy but is crucial in particular for determining the scope of the decision problem to be analysed.[12] When we take artificial agents to be moral proxies for low-level human agents, it is natural to frame the relevant decisions to be made by artificial agents from the perspective of the low-level human agent. For instance, we could consider various problematic driving scenarios a self-driving car might find itself in, and then discuss how the car should confront these problems on behalf of the driver. Call this 'low-level agential framing'. When we take artificial agents to be moral proxies for high-level agents, on the other hand, we should frame the relevant decisions to be made by artificial agents from the perspective of those high-level agents. To use the example of self-driving cars again, from the perspective of designers or regulators, we should consider the aggregate consequences of many self-driving cars repeatedly confronting various problematic driving scenarios in accordance with their programming. Call this 'high-level agential framing'.

The issues of responsibility ascription and decision structuring are of course connected: when it is appropriate to frame a decision problem from the perspective of a particular agent, this is usually because the choice to be made falls under that agent's responsibility. Those who think of artificial agents as moral proxies for low-level agents often argue in favour of a greater degree of control on the part of individual users, for instance by having personalisable ethics settings, whereby the users can alter their artificial agent's programming to more closely match their own moral views.[13] Given such control, both decision structuring as well as most of the responsibility for the resulting choices should be low-level. But it is important to note here that the appropriate level of agential framing of the relevant decision problems and the level of agency at which we ascribe responsibility may in principle be different. We could, for instance, think of designers of

[11] Or, depending on your views on proper 'identity design' (see Conitzer, 'Designing Preferences' (n 10)) one single artificial agent making decisions in many decision contexts previously faced by many humans (e.g. a network of artificial trading agents acting in coordinated ways).

[12] See SO Hansson, 'Scopes, Options, and Horizons: Key Issues in Decision Structuring' (2018) 21 *Ethical Theory and Moral Practice* 259 for a very instructive discussion of this and other issues in decision structuring.

[13] See, e.g. A Sandberg and H Bradshaw-Martin, 'Autonomous Cars and their Moral Implications' (2015) 58(1) *Multitudes* 62; and G Contissa, F Lagioia, and G Sartor, 'The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law' (2017) 25 *Artificial Intelligence and Law* 365.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

artificial agents doing their best to design artificial agents to act on behalf of their users, but without giving the users any actual control over the design. As such, the designers could try to align the artificial agents with their best estimate of the users' considered and informed values. In that case, decision framing should be low-level. But insofar as low-level agents aren't actually in control of the programming of the artificial agents, we might think their responsibility for the resulting choices is diminished and should still lie mostly with the designers.

How should we respond to the Moral Proxy Problem for the purposes of decision structuring, then? In the literature on ethical dilemmas faced by artificial agents, a low-level response is often presupposed. The presumption of many authors there is that we can conclude fairly directly from moral judgements about individual dilemma situations (e.g., the much discussed trolley problem analogues) to how the artificial agents employed in the relevant context should handle them.[14] There is even an empirical ethics approach to making design decisions, whereby typical responses to ethical dilemmas that artificial agents might face are crowd-sourced, and then used to inform design choices.[15] This reflects an implied acceptance of artificial agents as low-level moral proxies. The authors mentioned who are arguing in favour of personalisable ethics settings for artificial agents also appear to be presupposing that the artificial agents they have in mind are moral proxies for low-level agents. The standard case for personalisable ethics settings is based on the idea that mandatory ethics settings would be unacceptably paternalistic. But imposing a certain choice on a person is only paternalistic if that choice was in the legitimate sphere of agency of that person in the first place. Saying that mandatory ethics settings are paternalistic thus presupposes that the artificial agents under discussion are moral proxies for low-level agents.

What could be a positive argument in favour of low-level agential framing? I can think of two main ones. The first draws on the debate over responsibility ascription. Suppose we thought that, in some specific context, the only plausible way of avoiding what are sometimes called 'responsibility gaps', that is, of avoiding cases where nobody can be held responsible for harms caused by artificial agents, was to hold low-level agents, and in particular users, responsible.[16] Now there seems to be something unfair about holding users responsible for choices by an artificial agent that (a) they had no design control over, and that (b) are only justifiable when framing the choices from a high-level agential perspective. Provided that, if we were to frame choices from a high-level agential perspective, we may sometimes end up with choices that are not justifiable from a low-level perspective, this provides us with an argument in favour of low-level agential framing. Crucially, however, this argument relies on the assumption that only low-level agents can plausibly be held responsible for the actions of artificial agents, which is of course contested, as well as on the assumption that there is sometimes a difference between what is morally justifiable when adopting a high-level and a low-level agential framing respectively, which I will return to.

[14] See, e.g. P Lin, 'Why Ethics Matter for Autonomous Cars' in M Maurer and others (eds), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte* (2015); G Keeling, 'Why Trolley Problems Matter for the Ethics of Automated Vehicles' (2020) 26 *Science and Engineering Ethics* 293 (hereafter Keeling, 'Trolley Problems'). See also J Himmelreich, 'Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations' (2018) 21 *Ethical Theory and Moral Practice* 669 (hereafter Himmelreich, 'Ethics of Autonomous Vehicles'); Nyholm and Smids, 'Ethics of Accident-Algorithms' (n 4); and A Jaques, 'Why the Moral Machine Is a Monster' (2019) University of Miami Law School: We Robot Conference (hereafter Jaques, 'Why the Moral Machine Is a Monster') who find issue with this.

[15] See E Awad and others, 'Crowdsourcing Moral Machines' (2020) 63(3) *Communications of the ACM* 48.

[16] On this and other solutions to the threat of responsibility gaps in the legal context see S Beck, 'The Problem of Ascribing Legal Responsibility in the Case of Robotics' (2016) 31 *AI Society* 473 (hereafter Beck, 'Ascribing Legal Personality'). For instance, German law assigns liability for damage caused by parking distance control systems to individual users.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

A second potential argument in favour of a low-level response to the Moral Proxy Problem is based on the ideal of liberal neutrality, which is indeed sometimes invoked to justify anti-paternalism of the form proponents of personalisable ethics settings are committed to. The moral trade-offs we can expect many artificial agents to face are often ones there is reasonable disagreement about. We are then, in Rawlsian terms, faced with a political, not a moral problem:[17] how do we ensure fair treatment of all given reasonable pluralism? In such contexts, one might think higher-level agents, such as policy-makers or tech companies should maintain liberal neutrality; they should not impose one particular view on an issue that reasonable people disagree on. One way of maintaining such neutrality in the face of a plurality of opinion is to partition the moral space so that individuals get to make certain decisions themselves.[18] In the case of artificial agents, such a partition of the moral space can be implemented, it seems, by use of personalisable ethics settings, which implies viewing artificial agents as moral proxies for low-level agents.

At the same time, we also find in the responses to arguments in favour of personalisable ethics settings some reasons to think that perhaps there is not really much of a conflict, in practice, between taking a high-level and a low-level agential perspective. For one, in many potential contexts of application of artificial agents, there are likely to be benefits from coordination between artificial agents that each individual user can in fact appreciate. For instance, *Jan Gogoll* and *Julian Müller* point out the potential for collective action problems when ethics settings in self-driving cars are personalisable: each may end up choosing a 'selfish' setting, even though everybody would prefer a situation where everybody chose a more 'altruistic' setting.[19] If that is so, it is in fact in the interest of everybody to agree to a mandatory 'altruistic' ethics setting. Another potentially more consequential collective action problem in the case of self-driving cars is the tragedy of the commons when it comes to limiting emissions, which could be resolved by mandatory programming for fuel-efficient driving. And *Jason Borenstein*, *Joseph Herkert*, and *Keith Miller* point out the advantages, in general, of a 'systems-level analysis', taking into account how different artificial agents interact with each other, as their interactions may make an important difference to outcomes.[20] For instance, a coordinated driving style between self-driving cars may help prevent traffic jams and thus benefit everybody.

What this points to is that in cases where the outcomes of the choices of one artificial agent depend on what everybody else does and vice versa, and there are potential benefits for each from coordination and cooperation, it may seem like there will not be much difference between taking a low-level and a high-level agential perspective. From a low-level perspective, it makes sense to agree to not simply decide oneself how one would like one's artificial agent to choose. Rather, it is reasonable from a low-level perspective to endorse a coordinated choice where designers or regulators select a standardised programming that is preferable for each individual compared to the outcome of uncoordinated choice. And notably, this move does not need to be in tension with the ideal of liberal neutrality either: in fact, finding common principles that can be endorsed from each reasonable perspective is another classic way to ensure liberal neutrality

[17] As also pointed out by Himmelreich 'Ethics of Autonomous Vehicles' (n 14) and I Gabriel, 'Artificial Intelligence, Values and Alignment' (2020) 30 *Minds and Machines* 411.

[18] See J Gogoll and J Müller, 'Autonomous Cars: In Favor of a Mandatory Ethics Setting' (2017) 23 *Science and Engineering Ethics* 681 (hereafter Gogoll and Müller, 'Autonomous Cars') for this proposal, though they ultimately reject it. The phrase 'partition of the moral space' is due to G Gaus, 'Recognized Rights as Devices of Public Reason', in J Hawthorne (ed), *Ethics, Philosophical Perspectives* (2009), 119.

[19] Ibid.

[20] J Borenstein, J Herkert and K Miller, 'Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis' (2019) 25 *Science and Engineering Ethics* 383. See also Jaques, 'Why the Moral Machine Is a Monster' (n 14).

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

in the face of reasonable pluralism, in cases where partitioning the moral space in the way previously suggested can be expected to be worse for all. In the end, the outcome may not be so different from what a benevolent or democratically constrained high-level agent would have chosen if we thought of the artificial agents in question as high-level proxies in the first place.

Another potential reason for thinking that there may not really be much of a conflict between taking a high-level and a low-level agential perspective appears plausible in the remaining class of cases where we don't expect there to be much of an interaction between the choices of one artificial agent and any others. And that is simply the thought that in such cases, what a morally reasonable response to some choice scenario is should not depend on agential perspective. For instance, one might think that what a morally reasonable response to some trolley-like choice scenario is should not depend on whether we think of it from the perspective of a single low-level agent, or as part of a large number of similar cases a high-level agent is deciding on.[21] And if that is so, at least for the purposes of decision structuring, it would not make a difference whether we adopt a high-level or a low-level agential perspective. Moreover, the first argument we just gave in favour of low-level agential framing would be undercut.

Of course, while this may result in the Moral Proxy Problem being unimportant for the purposes of decision structuring, this does not solve the question of responsibility ascription. Resolving that question is not my primary focus here. What I would like to point out, however, is that the idea that agential framing is irrelevant for practical purposes sits nicely with a popular view on the question of responsibility ascription, namely the view that responsibility is often distributed among a variety of agents, including both high-level and low-level agents. Take, for instance, *Mariarosaria Taddeo* and *Luciano Floridi*:

> The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware [...] With distributed agency comes distributed responsibility.[22]

Shared responsibility between, amongst others, designers and users is also part of Rule 1 of 'the Rules' for moral responsibility of computing artefacts championed by *Miller*.[23] The reason why the idea of shared responsibility sits nicely with the claim that agential framing is ultimately practically irrelevant is that in that case, no agent can be absolved from responsibility on the grounds that whatever design choice was made was not justifiable from their agential perspective.The following discussion will put pressure on this position. It will show that in the context of risk, quite generally, agential perspective in decision structuring is practically relevant. This is problematic for the view that responsibility for the choices of artificial agents is often shared between high-level and low-level agents and puts renewed pressure on us to address the Moral Proxy Problem in a principled way. I will return to the Moral Proxy Problem in Section V to discuss why this is, in fact, a hard problem to address. In particular, it will become apparent that

[21] A Wolkenstein, 'What has the Trolley Dilemma Ever Done for Us (and What Will it Do in the Future)? On some Recent Debates about the Ethics of Self-Driving Cars' (2018) 20 *Ethics and Information Technology* 163 seems to make essentially this claim in response to criticism of the importance of trolley cases when thinking of the ethics of self-driving cars.

[22] M Taddeo and L Floridi, 'How AI Can Be a Force for Good' (2018) 361 *Science* 751, 751. See also M Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 *Science and Engineering* 2051. In the legal literature, there has been appeal to the idea of a legal 'electronic person' composed of designers, producers, users, etc. as a potential responsibility-bearer, see for example Beck, 'Ascribing Legal Personality' (n 16).

[23] K Miller, 'Moral Responsibility for Computing Artifacts: "The Rules"' (2011) 13(3) *IT Professional* 57.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

the low-level response to the problem that is so commonly assumed comes with significant costs in many applications. The high-level alternative, however, is not unproblematic either.

III. THE LOW-LEVEL CHALLENGE TO RISK NEUTRALITY IN ARTIFICIAL AGENT DESIGN

I now turn to the question of how to design artificial agents to deal with risk, which I will go on to argue is a practical design issue which crucially depends on our response to the Moral Proxy Problem. Expected utility theory is the orthodox theory of rational choice under conditions of risk that, on the standard approach, designers of artificial agents eventually aim to implement. The theory is indeed also accepted by many social scientists and philosophers as a theory of instrumentally rational choice, and moreover incorporated by many moral philosophers when theorising about our moral obligations in the context of risk.[24] Informally, according to this theory, for any agent, we can assign both a probability and a utility value to each potential outcome of the choices open to them. We then calculate, for each potential choice, the probability-weighted sum of the utilities of the different potential outcomes of that choice. Agents should make a choice that maximises this probability-weighted sum.

One widely discussed worry about applying expected utility theory in the context of artificial agent design is that when risks of harm are imposed on others, the application of expected utility theory implies insensitivity to how ex ante risks are distributed among the affected individuals.[25] For instance, suppose that harm to one of two individuals is unavoidable, and we judge the outcomes where one or the other is harmed to be equally bad. Expected utility theory then appears unable to distinguish between letting the harm occur for certain for one of the individuals, and throwing a fair coin, which would give each an equal chance of being harmed. Yet the latter seems like an intuitively fairer course of action.

In the following, I would like to abstract away as much as possible from this problem, but rather engage with an independent concern regarding the use of expected utility theory when designing artificial agents that impose risks on others. And that is that, at least under the interpretation generally adopted for artificial agent design, the theory implies risk neutrality in the pursuit of goals and values, and rules out what we will call 'pure' risk aversion (or pure risk seeking), as I will explain in what follows. Roughly, risk aversion in the attainment of some good manifests in settling for an option with a lower expectation of that good because the range of potential outcomes is less spread out, and there is thus a lesser risk of ending up with bad outcomes. For instance, choosing a certain win of £100 over a 50% chance of £300 would be a paradigmatic example of risk aversion with regard to money. The expected monetary value of the 50% gamble is £150. Yet, to the risk averse agent, the certain win of £100 may be preferable because the option does not run the risk of ending up with nothing.

Expected utility theory can capture risk aversion through decreasing marginal utility in the good. When marginal utility is decreasing for a good, that means that, the more an agent already has of a good, the less additional utility is assigned to the next unit of the good. In our example, decreasing marginal utility may make it the case that the additional utility gained from receiving

[24] Indeed, as remarks by Keeling exemplify in the case of this debate, moral philosophers often assume that there can be a division of labour between them and decision theorists, whereby issues to do with risk and uncertainty are settled by decision theorists alone. For more see Keeling, 'Trolley Problems' (n 14). The issues discussed in the following illustrate just one way in which this assumption is mistaken.

[25] On the general issue of fair risk imposition, see the useful overview by M Hayenhjelm and J Wolff, 'The Moral Problem of Risk Imposition: A Survey of the Literature' (2012) 20 *European Journal of Philosophy* 26.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

£100 is larger than the additional utility gained from moving from £100 to £300. If that is so, then the risk averse preferences we just described can be accommodated within expected utility theory: the expected utility of a certain £100 will be higher than the expected utility of a 50% chance of £300 – even though the latter has higher expected monetary value.

Whether this allows us to capture all ordinary types of risk aversion depends in part on what we think utility is. According to what we might call a 'substantive' or 'realist' understanding of utility, utility is a cardinal measure of degrees of goal satisfaction or value. On that view, expected utility theory requires agents to maximise their expected degree of goal satisfaction, or expected value. And having decreasing marginal utility, on this view, means that the more one already has of a good, the less one values the next unit, or the less the next unit advances one's goals. On this interpretation, only agents who have decreasing marginal utility in that sense are permitted to be risk averse within expected utility theory. What is ruled out is being risk averse beyond what is explainable by the decreasing marginal value of a good. Formally, expected utility theory does not allow agents to be risk averse with regard to utility itself. On this interpretation, that means agents cannot be risk averse with regard to degrees of goal satisfaction, or value itself, which is what the above reference to 'pure' risk aversion is meant to capture. For instance, on this interpretation of utility, expected utility theory rules out that an agent is risk averse despite valuing each unit of a good equally.[26]

Importantly for us, such a substantive conception of utility seems to be widely presupposed both in the literature on the design of artificial agents, as well as by those moral philosophers who incorporate expected utility theory when thinking about moral choice under risk. In moral philosophy, expected utility maximisation is often equated with expected value maximisation, which, as we just noted, implies risk neutrality with regard to value itself.[27] When it comes to artificial agent design, speaking in very broad strokes, on the standard approach we start by specifying the goals the system should be designed to pursue in what is called the 'objective function' (or alternatively, the 'evaluation function', 'performance measure', or 'merit function'). For very simple systems, the objective function may simply specify one goal. For instance, we can imagine an artificial nutritional assistant whose purpose it is simply to maximise caloric intake. But in most applications, the objective function will specify several goals, as well how they are to be traded off. For instance, the objective function for a self-driving car will specify that it should reach its destination fast; use little fuel; avoid accidents and minimise harm in cases of unavoidable accident; and make any unavoidable trade-offs between these goals in a way that reflects their relative importance.

[26] On this and other interpretations of utility, see J Thoma, 'Decision Theory' in R Pettigrew and J Weisberg (eds), *The Open Handbook of Formal Epistemology* (2019). Note that there is a way of understanding utility that is popular amongst economists which does not have that implication. On what we may call the 'formal' or 'constructivist' interpretation, utility is merely whatever measure represents an agent's preferences, provided these preferences are consistent with the axioms of a representation theorem for the version of expected utility theory one is advocating. According to that understanding, what expected utility theory requires of agents is having consistent preferences, so that they are representable as expected utility maximising. And in that case, having decreasing marginal utility just expresses the fact that one is risk averse, because that must be a feature of the agent's utility function if we are to capture her as risk averse and expected utility maximising. Importantly, on this view, because utility is not assumed to be a cardinal measure of value itself, we can allow for the utility function to exhibit decreasing marginal utility in value or degrees of goal satisfaction, thus allowing for pure risk aversion.

[27] Specifically in the debate about the ethics of artificial agents, this assumption is made, for example by A Hevelke and J Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 *Science and Engineering Ethics* 619; N Goodall, 'Away from Trolley Problems and toward Risk Management' (2016) 30 *Applied Artificial Intelligence* 820; Gogoll and Müller, 'Autonomous Cars' (n 17); and Keeling, 'Trolley Problems' (n 14) among many others.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

After we have specified the objective function, the artificial agent should be either explicitly programmed or trained to maximise the expectation of that objective function.[28] Take, for instance, this definition of rationality from *Stuart Russell* and *Peter Norvig's* textbook on AI:

> For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.[29]

According to the authors, the goal of artificial agent design is to implement this notion of rationality as well as possible. But this just means implementing expected utility theory under a substantive understanding of the utility function as a performance measure capturing degrees of goal satisfaction.[30]

So, we have seen that, by assuming a substantive interpretation of utility as a cardinal measure of value or degrees of goal satisfaction, many moral philosophers and designers of artificial agents are committed to risk neutrality with regard to value or goal satisfaction itself. However, such risk neutrality is not a self-evident requirement of rationality and/or morality. Indeed, some moral philosophers have defended a requirement to be risk averse, for instance when defending precautionary principles of various forms, or famously *John Rawls* in his treatment of choice behind the veil of ignorance.[31] And the risk neutrality of expected utility theory under the substantive interpretation of utility has been under attack recently in decision theory as well, for example by *Lara Buchak*.[32]

To illustrate, let me introduce two scenarios that an artificial agent might find itself in, where the risk neutral choice appears intuitively neither morally nor rationally required, and where indeed many human agents can be expected to choose in a risk averse manner.

Case 1: Artificial Rescue Coordination Centre. An artifical rescue coordination centre has to decide between sending a rescue team to one of two fatal accidents involving several victims. If it chooses Accident 1, one person will be saved for certain. If it chooses Accident 2, on the other hand, there is a 50% chance of saving three and a 50% chance of saving nobody. It seems plausible in this case that the objective function should be linear in lives saved, all other things being equal – capturing the idea that all lives are equally valuable. And let us suppose that all other morally

[28] On the difference between top-down and bottom-up approaches to implementing ethical design, see C Allen, I Smit, and W Wallach, 'Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches' (2005) 7 *Ethics and Information Technology* 149. What is important for us is the ideal of maximising the expectation of the objective function that is shared by both.

[29] S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., 2020) 37.

[30] It should be noted here that these authors also speak explicitly of 'utility' as distinct from the performance measure, because they think of the utility function more narrowly as something that is used internally by an agent to compute an optimal choice. For those agents that are programmed to be such explicit expected utility maximisers, the authors do remark that 'an agent's utility function is essentially an internalization of the performance measure. If the internal utility function and the external performance measure are in agreement, then an agent that chooses actions to maximize its utility will be rational according to the external performance measure.' (Ibid, 53) But note that expected utility theory is not normally understood to require explicit deliberation involving a utility function, but to also accommodate agents whose choices de facto maximise expected utility, no matter what deliberative procedure they use. Russell and Norvig's definition of rationality captures this wider conception of expected utility theory if we think of utility and performance measure as equivalent.

[31] J Rawls, *A Theory of Justice* (1971).

[32] L Buchak, *Risk and Rationality* (2013) (hereafter Buchak, *Risk and Rationality*).

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

relevant factors are indeed equal between the two options open to the rescue coordination centre.[33] In this scenario, a risk neutral rescue coordination centre would always choose Accident 2, because the expected number of lives saved (1.5) is higher. However, I submit that many human agents, if they were placed in this situation with time to deliberate, would choose Accident 1 and thus exhibit risk aversion. Moreover, doing so is neither intuitively irrational nor immoral. If this is not compelling, consider the case where attending to Accident 2 comes with only a 34% chance of saving three. Risk neutrality still requires choosing Accident 2. But it is very hard to see what would be morally or rationally wrong with attending to Accident 1 and saving one for certain instead.

Case 2: Changing Lanes. A self-driving car is driving in the left lane of a dual carriageway in the UK and is approaching a stumbling person on the side of the road. At the same time, a car with two passengers is approaching from behind on the right lane. The self-driving car estimates there is a small chance the other car is approaching fast enough to fatally crash into it should it change lanes (Changing Lanes), and a small albeit three times higher chance that the person on the side of the road could trip at the wrong time and consequently be fatally hit by the self-driving car should it not change lanes (Not Changing Lanes). Specifically, suppose that Not Changing Lanes comes with a 0.3% chance of killing one, meaning the expected number of fatalities is 0.003. Changing Lanes, on the other hand, comes with a 0.1% chance of killing two, meaning the expected number of fatalities is 0.002. Suppose that the passenger of the self-driving car will be safe either way. It seems plausible that the objective function should be linear in accidental killings, all other things being equal – again capturing the idea that all lives are equally valuable. And let us again suppose that all other morally relevant factors are indeed equal between the two options open to the self-driving car. In this scenario, a risk neutral car would always choose Changing Lanes, because the expected number of fatalities is lower. However, I submit that many human agents would, even with time to reflect, choose Not Changing Lanes to rule out the possibility of killing 2, and thus exhibit risk aversion. Moreover, doing so is neither intuitively irrational nor immoral.

These admittedly very stylised cases were chosen because they feature an objective function uncontroversially linear in the one value at stake, in order to illustrate the intuitive permissibility of pure risk aversion. Most applications will, of course, feature more complex objective functions trading off various concerns. In such cases, too, the standard approach to artificial agent design requires risk neutrality with regard to the objective function itself. But again, it is not clear why risk aversion should be ruled out, for example when a nursebot that takes into account both the potential value of saving a life and the cost of calling a human nurse faces a risky choice about whether to raise an alarm.

In the case of human agents, we tend to be permissive of a range of pure risk attitudes, including different levels of pure risk aversion. There appears to be rational and moral leeway on degrees of risk aversion, and thus room for reasonable disagreement. Alternatives to expected utility theory, such as Buchak's risk-weighted expected utility theory, as well as expected utility theory under some interpretations other than the substantive one, can accommodate such

[33] Note in particular that, if we don't allow randomisation between the two options, ex ante equality is impossible to achieve, and thus not a relevant factor.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

rational leeway on attitudes towards risk.[34] But the commitment to expected utility theory under a substantive interpretation of utility, as we find it in the literature on the design of artificial agents, rules this out and imposes risk neutrality instead – which is a point not often acknowledged, and worth emphasising.

To return to the Moral Proxy Problem, suppose that we want artificial agents to be low-level moral proxies. In the preceding examples, we have already picked the right agential framing then: we have structured the relevant decision problem as an individual choice situation as it might previously have been faced by a low-level human agent. A low-level moral proxy should, in some relevant sense, choose in such a way as to implement considered human judgement from the low-level perspective. Under risk, this plausibly implies that we should attempt to align not only the artificial agent's evaluations of outcomes, but also its treatment of risk to the values and attitudes of the low-level agents it is a moral proxy for. There are different ways of making sense of this idea, but on any such way, it seems like we need to allow for artificial agents to sometimes exhibit risk aversion in low-level choices like the ones just discussed.

As we have seen before, some authors who view artificial agents as low-level moral proxies have argued in favour of personalisable ethics settings. If there is, as we argued, reasonable disagreement about risk attitudes, artificial agents should then also come with personalisable risk settings. If we take an empirical approach and crowd-source and then implement typical judgements on ethical dilemma situations like the ones just discussed, we will likely sometimes need to implement risk averse judgements as well. Lastly, in the absence of personalisable ethics and risk settings but while maintaining the view of artificial agents as low-level moral proxies, we can also view the design decision as the problem of how to make risky choices on behalf of another agent while ignorant of their risk attitudes. One attractive principle for how to do so is to implement the most risk averse of the reasonable attitudes towards risk, thereby erring on the side of being safe rather than sorry when choosing for another person.[35] Again, the consequence would be designing artificial agents that are risk averse in low-level decisions like the ones we just considered.

We have seen, then, that in conflict with the standard approach to risk in artificial agent design, if we take artificial agents to be low-level moral proxies, we need to allow for them to display pure risk aversion in some low-level choice contexts like the ones just considered. The next section will argue that things look quite different, however, if we take artificial agents to be high-level moral proxies.

IV. RISK AVERSION AND THE HIGH-LEVEL AGENTIAL PERSPECTIVE

Less stylised versions of the scenarios we just looked at are currently faced repeatedly by different human agents and will in the future be faced repeatedly by artificial agents. While such decisions are still made by human agents, there is usually nobody who is in control of a large number such choice problems: human rescue coordinators will usually not face such a dramatic decision multiple times in their lives. And most drivers will not find themselves in such dangerous driving situations often. The regulatory reach of higher-order agents such as policymakers over human agents is also likely to be limited in these scenarios and many other areas in which artificial agents might be introduced to make decisions in place of humans – both because human agents in such choice situations have little time to reflect and will thus often

[34] Buchak, *Risk and Rationality* (n 32).

[35] For a defence of such a principle see L Buchak, 'Taking Risks Behind the Veil of Ignorance' (2017) 127(3) *Ethics* 610.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

be excused for not following guidelines, and because, in the case of driving decisions in particular, there are limits to the extent to which drivers would accept being micromanaged by the state.

Things are different, however, once artificial agents are introduced. Now there are higher-level agents, in particular designers, who can directly control the choice behaviour of many artificial agents in many instances of the decision problems we looked at in the last section. Moreover, these designers have time to reflect on how decisions are to be made in these choice scenarios and have to be explicit about their design choice. This also gives greater room for other higher-level agents, such as policy-makers, to exert indirect control over the choices of artificial agents, by regulating the design of artificial agents. Suppose we think that artificial agents in some specific context should in fact be thought of as moral proxies not for low-level agents such as individual users of self-driving cars, but rather as moral proxies for such high-level agents. From the perspective of these higher-level agents, what seems most relevant for the design choice are the expected aggregate consequences of designing a whole range of artificial agents to choose in the specified ways on many different occasions. I want to show here that this makes an important difference in the context of risk.

To illustrate, let us return to our stylised examples, starting with a modified version of Case 1: Artificial Rescue Coordination Centre:

> Suppose some high-level agent has to settle at once on one hundred instances of the choice between Accident 1 and Accident 2. Further, suppose these instances are probabilistically independent, and that the same choice needs to be implemented in each case. The two options are thus always going for Accident 1, saving one person for certain each time, or always going for Accident 2, with a 50% chance of saving three each time. The expected aggregate outcome of going for Accident 1 one hundred times is, of course, saving one hundred people for certain. The expected aggregate result of going for Accident 2 one hundred times, on the other hand, is a probability distribution with an expected number of one hundred and fifty lives saved, and, importantly, a <0.5% chance of saving fewer lives than if one always went for Accident 1. In this compound case, it now seems unreasonably risk averse to choose the 'safe option'.

Similarly, if we look at a compound version of Case 2: Changing Lanes:

> Suppose a higher-level agent has to settle at once how 100,000 instances of that choice should be made, where these are again assumed to be probabilistically independent, and the same choice has to be made on each instance. One could either always go for the 'safe' option of Not Changing Lanes. In that case, the expected number of fatalities is 300, with a <0.1% chance of less than 250 fatalities. Or one could always go for the 'risky' option of Changing Lanes. In that case, the expected number of fatalities is only 200, with only a ~0.7% chance of more than 250 fatalities. As before, the 'risky' option is thus virtually certain to bring about a better outcome in the aggregate, and it would appear unreasonably risk averse to stick with the 'safe' option.

In both cases, as the number of repetitions increases, the appeal of the 'risky' option only increases, because the probability of doing worse than on the 'safe' option becomes ever smaller. We can also construct analogous examples featuring more complex objective functions appropriate for more realistic cases. It remains true that as independent instances of the risky choice problem are repeated, at some point the likelihood of doing better by each time choosing a safer option with lower expected value becomes very small. From a sufficiently large compound perspective, the virtual certainty of doing better by picking a riskier option with higher expected value is decisive. And thus, when we think of artificial agents as moral proxies for high-level agents that are in a position to control sufficiently many low-level decisions, designing the

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

artificial agents to be substantially risk averse in low-level choices seems impermissible. From the high-level agential perspective, the risk neutrality implied by the current standard approach in artificial agent design seems to, in fact, be called for.[36]

The choice scenarios we looked at are similar to a case introduced by *Paul Samuelson*,[37] which I discuss in more detail in another paper.[38] *Samuelson's* main concern there is that being moderately risk averse in some individual choice contexts by, for example, choosing the safer Accident 1 or Not Changing Lanes, while at the same time choosing the 'risky' option in compound cases is not easily reconcilable with expected utility theory (under any interpretation).[39] It is undeniable, though, that such preference patterns are very common. And importantly, in the cases we are interested in here, no type of agent can actually be accused of inconsistency, because we are dealing with two types of agents with two types of associated choice problems. One type of agent, the low-level agent who is never faced with the compound choice, exhibits the reasonable seeming risk averse preferences regarding 'small-scale' choices to be made on her behalf. And another type of agent, the high-level agent, exhibits again reasonable-seeming preferences in compound choices that translate to effective risk neutrality in each individual 'small-scale' choice scenario.

The take-away is thus that how we respond to the Moral Proxy Problem is of practical relevance here: If we take artificial agents to be moral proxies for low-level agents, they will sometimes need to be programmed to exhibit risk aversion in the kinds of individual choice contexts where they are replacing human agents. If we take them to be moral proxies for high-level agents, they should be programmed to be risk neutral in such choice contexts, as the current approach to risk in artificial agent design in fact implies, because this has almost certainly better consequences in the aggregate.

V. BACK TO THE MORAL PROXY PROBLEM

We saw in Section II that the Moral Proxy Problem matters for decision structuring: whether we take artificial agents to be moral proxies for low-level or high-level agents determines from which agential perspective we are framing the relevant decision problems. I raised the possibility, alluded to by some authors, that resolving the Moral Proxy Problem one way or the other is of little practical relevance, because agential framing does not make a practical difference for design choices. The issue of whether artificial agents should be designed to be risk neutral or allowed to be risk averse, discussed in the last two sections, is then an especially challenging one in the context of the Moral Proxy Problem, because it shows the hope for this irrelevance to be ungrounded: agential perspective turns out to be practically crucial.

Notably, the stylised examples we discussed do not describe collective action or coordination problems where each can recognise from her low-level perspective that a higher-level agent could implement a coordinated response that would be superior from her perspective and everybody else's. Crucially, both the outcomes and the probabilities in each of the lower-level

[36] To the extent that even low-level agents face some risky decisions very often, we may also take this to be an argument that in those cases, risk neutrality in the individual choice instances is called for even from the low-level perspective. However, in our examples, the individual choice scenarios are both rare and high-stakes from the low-level perspective, so that the compound perspective really only becomes relevant for high-level agents. It is in that kind of context that agential perspective makes a crucial practical difference.

[37] P Samuelson, 'Risk and Uncertainty: A Fallacy of Large Numbers' (1963) 98 *Scientia* 108 (hereafter Samuelson, 'Risk and Uncertainty').

[38] J Thoma, 'Risk Aversion and the Long Run' (2019) 129(2) *Ethics* 230.

[39] Samuelson, 'Risk and Uncertainty' (n 37).

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

choice contexts are independent in our examples. And having a particular design imposed by a higher-level agent does not change the potential outcomes and probabilities of the choice problem faced by any particular artificial agent. It only changes the actual choice in that lower-level choice problem from a potentially risk averse to a risk neutral one. This is not something that a risk averse lower-level agent would endorse.

It thus becomes practically important to resolve the Moral Proxy Problem. And for the purposes of decision structuring, at least, it is not an option to appeal to the notion of distributed agency to claim that artificial agents are moral proxies for both low-level and high-level agents. Adopting one or the other agential perspective will sometimes call for different ways of framing the relevant decision problem, and we need to settle on one specification of the decision problem before we can address it. Where we imagine there being a negotiation between different stakeholders in order to arrive at a mutually agreeable result, the framing of the decision problem to be negotiated on will also need to be settled first. For decision structuring, at least, we need to settle on one agential perspective.

For reasons already alluded to, the fact that substantially different designs may be morally justified when decision problems are framed from the high-level or the low-level agential perspective is also problematic for ascribing shared responsibility for the choices made by artificial agents. If different programmings are plausible from the high-level and low-level perspective, it may seem unfair to hold high-level agents (partially) responsible for choices justified from the low-level perspective and vice versa. If, based on a low-level framing, we end up with a range of risk averse self-driving cars that cause almost certainly more deaths in the aggregate, there is something unfair about holding designers responsible for that aggregate result. And if, based on a high-level framing, we in turn end up with a range of risk neutral self-driving cars, which, in crash scenarios frequently save nobody when they could have saved some for sure, there is something unfair about holding individual users responsible for that tough call they would not have endorsed from their perspective.[40] At least, it seems like any agent who will be held responsible for some (set of) choices has some rightful claim for the decision problem to be framed from their agential perspective. But where agential perspective makes a practical difference not all such claims can be fulfilled.

Let us return now to the problem of decision structuring, where, for the reasons just mentioned, we certainly need to resolve the Moral Proxy Problem one way or the other. However we resolve it, there are major trade-offs involved. I already mentioned some potential arguments in favour of low-level agential framing. There is, for one, the idea that low-level agential framing is natural if we want to hold low-level agents responsible. If we don't have an interest in holding low-level agents responsible, this is, of course, not a relevant consideration. But I would also like to add an observation about moral phenomenology that may have at least some political relevance. Note that users and owners of artificial agents are in various senses morally closer to the potentially harmful effects of the actions of their artificial agents than designers or policy-makers: they make the final decision of whether to deploy the agent; their lives may also be at stake; they often more closely observe the potentially harmful event and have to live with its memory; and users are often asked to generally maintain responsible oversight of the operations of the artificial agent. All this may, at least, result in them feeling more responsible for the actions of their artificial agent. Such a

[40] Granted, individual users usually do make the final call of whether to deploy an artificial agent and may do so knowing how they would act in certain morally difficult situations. Still, if certain aspects of the programming of the artificial agent one deploys only make sense from the perspective of general public safety, or general public health, and only in the context of many other artificial agents being programmed in the same way, it is natural to resist individual responsibility for the consequences of that aspect of the artificial agent's design.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

feeling of responsibility and moral closeness without control, or without at least the sense that efforts were made for the choices of the artificial agent to capture one's considered judgements as well as possible is a considerable burden.

A second argument we made in favour of low-level agential framing appealed to the idea of liberal neutrality in the face of reasonable disagreement, which could be implemented effectively by partitioning the moral space so as to leave certain decisions up to individuals. Such partitioning seems like an effective way to implement liberal neutrality especially in the absence of collective action problems that may create general agreement on a coordinated response. Given the independence in outcomes and probabilities, the cases we have discussed indeed do not constitute such collective action problems, but they do feature reasonable disagreement in the face of rational and moral leeway about risk attitudes. I believe that the ideal of liberal neutrality is thus a promising consideration in favour of low-level agential framing.

What the preceding sections have also made clear, however, is that low-level agential framing in the context of risk may come at the cost of aggregate outcomes that are almost certainly worse than the expected consequences of the choices that seem reasonable from the high-level agential perspective. This consequence of low-level agential framing is, as far as I know, unacknowledged, and may be difficult for proponents of low-level agential framing to accept.

If we respond to the Moral Proxy Problem by adopting a high-level agential perspective in those contexts instead, this problem is avoided. And other considerations speak in favour of thinking of artificial agents as moral proxies for high-level agents. An intuitive thought is this: as a matter of fact, decisions that programmers and those regulating them make determine many lower-level choices. In that sense they are facing the compound choice, in which the almost certainly worse aggregate outcome of allowing lower-level risk aversion appears decisive. In order to design artificial agents to be (risk averse) moral proxies for individual users, designers would have to abstract away from these very real aggregate implications of their design decisions. This may put designers in a similarly difficult position to the owner of a self-driving car that she knows may make choices that seem reckless from her perspective.

Following on from this, arguments in favour of holding high-level agents responsible will also, at least to some extent, speak in favour of high-level agential framing, because again it seems high-level agential framing is natural when we want to hold high-level agents responsible. We find one potential argument in favour of ascribing responsibility to high-level agents in *Hevelke* and *Nida-Rümelin's* appeal to moral luck.[41] Their starting point is that whether individual artificial agents ever find themselves in situations where they have to cause harm is in part down to luck. For instance, it is in part a matter of luck whether, and if so how often, any artificial agent finds itself in a dangerous driving situation akin to the one described in Case 2 mentioned earlier. And, no matter how the agent chooses, it is a further matter of luck whether harm is actually caused. Where harm is caused, it may seem unfair to hold the unlucky users of those cars responsible, but not others who employed their artificial agents no differently. *Alexander Hevelke* and *Julian Nida-Rümelin* take this observation to speak in favour of ascribing responsibility collectively to the group of all users of a type of artificial agent. But finding responsibility with other high-level agents, such as the companies selling the artificial agents would also avoid the problem of moral luck. And then it also makes sense to adopt a high-level perspective for the purposes of decision structuring.

[41] A Hevelke and J Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 *Science and Engineering Ethics* 619.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press

Still, the practical relevance of agential framing also brings about and highlights costs of settling for a high-level solution to the Moral Proxy Problem that are worth stressing: this solution will mean that, where artificial agents operate in areas where previously human agents made choices, these artificial agents will make some choices that are at odds with even considered human judgement. And the high-level solution will introduce higher-level control, be it by governments or tech companies, in areas where previously decision-making by humans has been decentralised, and in ways that don't simply reproduce what individual human agents would have (ideally) chosen for themselves. In this sense, the high-level solution involves a significant restructuring of our moral landscape.

VI. CONCLUSION

I have argued that the Moral Proxy Problem, the problem of determining what level of human (group) agent artificial agents ought to be moral proxies for, has special practical relevance in the context of risk. Moral proxies for low-level agents may need to be risk averse in the individual choices they face. Moral proxies for high-level agents, on the other hand, should be risk neutral in individual choices, because this has almost certainly better outcomes in the aggregate. This has a number of important implications. For one, it means we actually need to settle, in any given context, on one response to the Moral Proxy Problem for purposes of decision structuring at least, as we don't get the same recommendations under different agential frames. This, in turn, puts pressure on the position that responsibility for the choices of artificial agents is shared between high-level and low-level agents.

My discussion has also shown that any resolution of the Moral Proxy Problem involves sacrifices: adopting the low-level perspective implies designers should make design decisions that have almost certainly worse aggregate outcomes than other available design decisions, and regulators should not step in to change this. Adopting the high-level perspective, on the other hand, involves designers or regulators imposing specific courses of action in matters where there is intuitively rational and moral leeway when human agents are involved and where, prior to the introduction of new technology, the state and tech companies exerted no such control. It also risks absolving users of artificial agents of felt or actual responsibility for the artificial agents they employ, and having them live with consequences of choices they would not have made.

Finally, I have shown that because the way in which expected utility theory is commonly understood and implemented in artificial agent design implies risk neutrality regarding goal satisfaction, it involves, in a sense, a tacit endorsement of the high-level response to the Moral Proxy Problem which makes such risk neutrality generally plausible. Given low-level agential framing, risk aversion is intuitively neither always irrational nor immoral, and is in fact common in human agents. The implication is that if we prefer a low-level response to the Moral Proxy Problem in at least some contexts, risk aversion should be made room for in the design of artificial agents. Whichever solution to the Moral Proxy Problem we settle on, I hope my discussion has at least shown that the largely unquestioned implementation of risk neutrality in the design of artificial agents deserves critical scrutiny and that such scrutiny reveals that the right treatment of risk is intimately connected with how we answer difficult questions about agential perspective and responsibility in a world increasingly populated by artificial agents.

https://doi.org/10.1017/9781009207898.006 Published online by Cambridge University Press