



Representative Robots: Can AI Systems Act in Our Name?

Isaac Taylor, Department of Philosophy, Stockholm University,
isaac.taylor@philosophy.su.se

Using AI systems to make decisions in the place of humans promises greater efficiency, but some authors raise a number of ethical worries about this. The undermining of responsibility, the removal of humans from meaningful participation, and a misalignment of values, it has been claimed, may result from a reliance on algorithmic decision-making. Most ways of avoiding these problems that have been proposed involve somehow ensuring that humans have direct control over the AI systems in question. But there is reason to think that there are limits to what this strategy can achieve. In this paper, I propose an alternative strategy. Building AI systems in a way that allows them to act in the name of humans, I suggest, would allow us to avoid the problems without the need for direct control. Drawing on recent work in democratic theory, I then claim that it is possible, in principle, for AI systems to act in our name. Achieving the benefits of AI without the costs that are often associated with their use may depend on our ability to create AI systems that function as our representatives.



Representative Robots: Can AI Systems Act in Our Name?

Isaac Taylor

I. INTRODUCTION

Could an AI system act in our name? This is the question I want to consider. On the face of it, a paper addressing this question might strike some as among the worst excesses of philosophical speculation. But I beg the reader's patience. If it were the case that we could (perhaps by placing limits on the development and use of AI) have AI systems that are proper representatives of us, this may allow us to mitigate some of the potential harms of high-stakes AI use that have been identified in recent work. As we will see, many purported moral costs of AI use stem from the fact that humans cannot adequately control the systems in question or predict how they will behave once deployed. Yet, if the systems can nonetheless be said to act in the name of humans, we may be able to avoid the costs without having the sort of direct control or foresight that is sometimes thought to be necessary. I discuss some of these costs, and why they might be avoided when AI systems that act in our name are used, in Section II.

In order to see whether AI systems can act in our name, we need to consider what conditions need to be in place for one agent to act in the name of another. We then need to investigate whether AI systems can meet these conditions. I undertake these tasks in Sections III, IV, and V. Section III looks at what I call "external conditions," which relate to the observable behavior of purported representatives. Section IV looks at "internal conditions," relating to their mental states (reasons, intentions, and the like). Section V, finally, looks at "relational conditions," which concern the necessary interactions and relations between purported representatives and those they seek to represent.

I extract the conditions from recent work in democratic theory. Different theorists have put forward different combinations of conditions that they think, in combination, give rise to genuine representation. But, for clarity, instead of discussing theories of representation I will simply discuss the conditions that make up different theories sequentially. This helps us avoid a worry that there are different forms of representation, each with its own constitutive and normative criteria, as each form

may involve a combination of different conditions.¹ Suppose, for instance, that we want to distinguish between anticipatory representation, in which individuals have power over their representative through their ability to vote them out of office, and surrogate representation, which occurs when representatives represent those outside their constituency of voters.² To become a representative on the former model, a formal ability to get rid of representatives on the part of those represented is required. No such ability is needed on the latter model, although this might give rise to more demanding requirements of deliberation.³ If we find that AI systems can meet all the conditions of which different forms of representation require sub-sets, we can be confident that AI systems can act in our name in a variety of ways.

I will ultimately argue that AI systems can in principle meet all the conditions examined (or at least justifiably modified versions of them). The perhaps surprising answer I will give to my question, therefore, is “yes” (assuming some combination of conditions examined is sufficient for representation). It is a qualified “yes,” however, since our ability to ensure that the conditions of representative agency are met by AI systems may vary by context. I am interested in this paper solely with whether it is possible in principle to build (existing or near-future⁴) AI systems that can act in our name. How often we will be able to do this in practice is a question for future research. I return to this point in the conclusion.

II. WHY REPRESENTATION?

Why might we think that trying to build AI systems that act in our name is a good idea in the first place? Consider the following three challenges, well-known in the field of AI ethics.

First, it is often thought that deploying certain sorts of AI systems to complete certain sorts of tasks will generate a problematic “responsibility gap,” that is, a situation in

¹ Jane Mansbridge, “Rethinking Representation,” *American Political Science Review* 97, no. 4 (2003): 515–28, <https://doi.org/10.1017/S0003055403000856>; Jane Mansbridge, “Clarifying the Concept of Representation,” *American Political Science Review* 105, no. 3 (2011): 621–39, <https://doi.org/10.1017/S0003055411000189>; Andrew Rehfeld, “Representation Rethought: On Trustees, Delegates, and Gyroscopes in the Study of Political Representation and Democracy,” *American Political Science Review* 103, no. 2 (2009): 214–30, <https://doi.org/10.1017/S0003055409090261>; Andrew Rehfeld, “The Concepts of Representation,” *American Political Science Review* 105, no. 3 (2011): 631–41, <https://doi.org/10.1017/S0003055411000190>.

² Mansbridge, “Rethinking Representation,” 516–20, 522–25.

³ *Ibid.*, 525.

⁴ I am not interested, for instance, if an advanced system that was a full agent could act in our name – the answer would presumably be yes (so long as we think that other full agents, such as humans, can act in our name).

which nobody is responsible for the actions taken.⁵ Worries about a responsibility gap emerging have particularly arisen in discussions about autonomous weapons systems (AWSs), which would be able to select and engage targets in an armed conflict without any direct human oversight. Robert Sparrow, in his influential paper on the subject, argues that “it is a fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies killed in the course of it.”⁶ And yet, says Sparrow, since programmers, military commanders, and the machines themselves cannot be (justly) held responsible for the actions of AWSs,⁷ there is a significant moral cost to deploying them.

Second, John Danaher has argued that the extensive use of algorithms in governmental decision-making generates a “threat of algocracy.”⁸ Just as we might be tempted to hand over power to an elite who are epistemically better positioned to make political decisions, there is a rationale for handing over control of some decisions to AI systems, which will have capacities that exceed humans in a number of areas. Yet, according to Danaher, doing so may “constrain the opportunities for human participation in, and comprehension of, public decision-making.”⁹

Third, AI ethicists are often concerned with the “value alignment problem.” In one notable statement of the idea, Stewart Russell explains that we might “perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own.”¹⁰ As the role AI systems play in our lives increases, even a small misalignment between the values that are decipherable in their programming and our own values may lead to ever greater set-backs to human interests and welfare.

Although these are very different challenges, they are generally all thought to arise from the same source, namely, the *autonomy* of certain AI systems. The term “autonomy” is used in different ways when talking about AI, but two senses of autonomous functioning are worth highlighting here. The first relates to the nature of

⁵ Andreas Mathias, “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,” *Ethics & Information Technology* 6, no. 3 (2004): 175–83, <https://doi.org/10.1007/s10676-004-3422-1>; Robert Sparrow, “Killer Robots,” *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77, <https://doi.org/10.1111/j.1468-5930.2007.00346.x>.

⁶ Ibid., 67.

⁷ Ibid., 69–73.

⁸ John Danaher (2016) “The Threat of Algocracy: Reality, Resistance and Accommodation,” *Philosophy & Technology* 29, 2016, 245–268, <https://doi.org/10.1007/s13347-015-0211-1>. John Danaher, “Freedom in an Age of Algocracy,” in *The Oxford Handbook of Philosophy of Technology*, ed. Shannon Vallor (Oxford: Oxford University Press, 2020), 250–72.

⁹ Ibid., 246.

¹⁰ Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (London: Allen Lane, 2019), 137.

the activity that AI systems might be tasked with: autonomous systems are defined as systems that can complete “higher-level” tasks, that is, tasks that are made of various lower-level sub-tasks and, crucially, can be completed in a number of different ways (by completing different combinations of sub-tasks, for example).¹¹ A simple machine that moves chess pieces around a board in line with my instructions completes only lower-level tasks, but a system that is designed to play and attempt to win at chess can complete this task by adopting a variety of different strategies.¹² This second sort of system, then, is more autonomous than the first in this sense of the term.

It would be relatively easy to design a chess-playing robot that required human authorization before completing each move. But sometimes such human oversight is impossible or undesirable. AWSs may often need to be deployed in battlefield settings where constant contact with a human operator cannot be guaranteed. And future generations of self-driving cars may need to make split-second decisions about what to do in evolving accident scenarios, and cannot wait for approval by a human passenger. Such systems would be autonomous in the second sense I have in mind here: they would act independently of direct human control.

Systems that exhibit both forms of autonomy simultaneously are taken by some AI ethicists to be particularly likely to give rise to the three challenges discussed above. Humans cannot predict how these systems will act in advance (owing to the first type of autonomy) or directly control them as they are acting (owing to the second type of autonomy). Consequently, those humans may not be responsible for the actions, they may not meaningfully participate in the decision-making, and the outcomes may not reflect their values.

It is no surprise, then, that solutions to these problems often take the form of ensuring that control can nonetheless be exercised by humans, through either technological or organizational interventions.¹³ The concept of “meaningful human control,” taken to be the gold standard of ethical AI in ongoing debates about AWSs,

¹¹ Giovanni Sartor and Andrea Omicini, “The Autonomy of Technological Systems and Responsibilities for Their Use,” in *Autonomous Weapons Systems: Law, Ethics, Policy*, eds. Nehal Bhuta et al. (Cambridge: Cambridge University Press, 2016), 48–51.

¹² Cf. Nick Bostrom & Elizer Yudkowsky, “The Ethics of Artificial Intelligence,” in *The Cambridge Handbook of Artificial Intelligence*, eds. Keith Frankish and William M. Ramsay (Cambridge: Cambridge University Press, 2014), 319.

¹³ Kevin Baum et al., “From Responsibility to Reason-Giving Explainable Artificial Intelligence,” *Philosophy & Technology* 35, no. 12 (2022): 1–30, <https://doi.org/10.1007/s13347-022-00510-w>; Danaher, “The Threat of Algocracy,” 258–65; Frank Hindriks and Herman Veluwenkamp, “The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps,” *Synthese* 201, no. 21 (2023): 11–17, <https://doi.org/10.1007/s11229-022-04001-5>.

captures both of these points.¹⁴ Without prejudging where these debates will go, we can simply note that limits have been identified to these sorts of solutions.¹⁵ It will thus be useful to consider if there are any alternatives.

My starting point in developing an alternative solution is the thought that we sometimes view ourselves responsible for the actions of other human beings, view ourselves as meaningfully participating in the decisions they make, and rest assured that their decisions will reflect our values even without direct control or the powers of prediction about how they will act. This is most often thought to be the case in well-functioning representative democracies, where our elected politicians are said to “act in our name.”

The concept of acting in the name of another is, I believe, an under-explored topic in philosophy. But the general idea appears to be fairly intuitive. It seems to occupy a space between two other sorts of action. On the one side, we have actions carried out by “proxy agents,” whereby the action of one agent simply *is* the action of another.¹⁶ On the other side, we have forms of action which are “owned” fully by the agent carrying them out, but which others may be more or less connected to in normatively significant ways (such as through inciting the action). When an action is done in our name, it cannot simply be described as ours since the agent carrying it out must exercise their own judgement to some extent (the details of the action to be carried out are not fully specified in advance).¹⁷ Yet, at the same time, we seem more intimately connected with the action than, say, if we merely incited, incentivized, or encouraged it. We seem to “own” the action in question. Successful acts of representation, to invoke Hanna Pitkin’s general definition of representation, ensure “the making present *in some sense* of something which is nevertheless *not* present literally or in fact,”¹⁸ namely the represented themselves.¹⁹

¹⁴ Filippo Santoni di Sio and Jeroen van den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Frontiers in Robotics and AI* 5, no. 15 (2018): 1–15, <https://doi.org/10.3389/frobt.2018.00015>.

¹⁵ Sparrow, “Killer Robots,” 68–69; Isaac Taylor, “Is Explainable AI Responsible AI?” *AI & Society* 40 (2025): 1695–1704, <https://doi.org/10.1007/s00146-024-01939-7>.

¹⁶ Kirk Ludwig, “Proxy Agency in Collective Action,” *Noûs* 48, no. 1 (2014): 75–105, <https://doi.org/10.1111/nous.12013>.

¹⁷ Contrast this with an archetypal case of proxy action: proxy voting. When someone casts a proxy vote for me, they must simply vote for my preferred candidate, even when they think that this is a bad choice (because I am not in full possession of all relevant facts about the candidate, for example).

¹⁸ Hanna Fenichel Pitkin, *The Concept of Representation* (Berkeley; Los Angeles: University of California Press, 1967).

¹⁹ There has been some recent work arguing that citizens are responsible for the actions of their state based on their participation or causal contribution to the state’s actions, for instance Eric Beerbohm, *In Our Name: The Ethics of Democracy* (Princeton: Princeton University Press, 2012); Avia Pasternak, *Responsible Citizens, Irresponsible States: Should Citizens Pay for Their State’s Wrongdoings?* (Oxford: Oxford University Press, 2024). For some doubts, see Holly Lawford-Smith, *Not in Their Name: Are Citizens Culpable for Their*

If AI systems can be thought of as acting in our name, we might be able to meet the challenges that their autonomy gives rise to. To see this, consider first the responsibility gap. It might be thought that another agent acting in our name renders us responsible for their actions.²⁰ If AI systems act in our name, then, there may be no responsibility gap, since we would be made responsible for the decisions taken.

As for the threat of algocracy, ensuring that public sector AI systems act in our name does not, of course, ensure that we literally participate directly in politics. But, if Pitkin's characterization is correct and representation involves making the people present in some sense, this may be sufficient to allay the worries associated with the slide into algocracy. It may be thought that the sort of simulation of presence characteristic of representation is enough. At least it will do no worse than familiar forms of representative democracy on this front. Unless we can introduce more radical forms of direct democracy into our government, representation by AI systems may be as good as the feasible alternative at ensuring participation (namely, representation by humans). One might think that the lack of direct participation in representative democracies can be compensated for by the influence that constituencies have over their representatives. However, as I show in Section V, something like this influence over autonomous AI systems is also possible.

Finally, consider how building representative AI systems might solve the value alignment problem. As we will see, some theorists think that acting in the name of another requires a sensitivity to reasons held by the other, or an openness to guidance by the other. If an AI system has one or both of these features, its values are unlikely to stray too far from those that it represents. Strictly speaking, it is not the fact that an AI system acts in the name of another that would solve the value alignment problem, but rather the fact that such a system meets various conditions that happen to be required to act in the name of another. Nonetheless, examining whether we can build representative AI systems is useful in considering this problem since, if we can, it is likely that values would be aligned as a side-effect without the need for control over the systems.

State's Actions? (Oxford: Oxford University Press, 2019). On such accounts, the actions of the state might be understood as the collective actions of the citizens. While it has been claimed, in a parallel manner, that the actions of AI systems are the collective actions of groups of enablers, in Michael Robillard, "No Such Thing as Killer Robots," *Journal of Applied Philosophy* 35, no. 4 (2018): 705–17, <https://doi.org/10.1111/ja12274>, there are objections to this view, put forward in Isaac Taylor, "Collective Responsibility and Artificial Intelligence," *Philosophy & Technology* 37, no. 27 (2024): 1–18, <https://doi.org/10.1007/s13347-024-00718-y>. Here, however, I am concerned with the distinct question of whether AI systems can be said to act in the name of others, which does not presuppose that the actions of those systems are literally the collective actions of groups of individuals.

²⁰ Avihay Dorfman and Alon Harel, *Reclaiming the Public* (Cambridge, Cambridge University Press, 2024), 15; Pitkin, *The Concept of Representation*, 18–19.

In closing this Section, it is worth addressing a worry that one might feel about the very concept of an artificial entity representing a human agent. Since our ideas of representation have primarily been developed in light of relationships among humans, is this not a dead-end from the start? Such concerns, I think, can begin to be addressed by taking note of a lesson from the work of Alan Turing. In his seminal paper on whether machines can think, Turing rejects the idea of investigating this question by beginning with an analysis of the concept of thinking that is in line with ordinary uses.²¹ I suspect this is, in part, because such a strategy would unduly bias the subsequent investigation by using an overly-anthropocentric notion of thought. Similarly, I suggest, we should be open to the possibility that acting in the name of another will look different if it is done by a machine rather than a human.

But perhaps the worry that some might be feeling concerns the very idea that machines can *act* at all (which seems to be a form of behavior exclusive to full agents). Although advanced AI systems seem to share some features with humans that are associated with agency, their lack of certain other features is thought by some to render them what we might call “quasi-agents.”²² While there might be something to this worry, we could allay it by simply reframing the investigation to be more about whether AI systems can perform a *functional equivalent* to acting in our name: that is, whether they can function in a way that is indistinguishable from humans acting in our name except for the fact that they are a different sort of entity. If we could establish this, I posit, the normative implications that follow from someone acting in our name will be identical to AI systems exhibiting the functional equivalent. With this point in mind, I will proceed by using the language of “acting in our name,” but the argument could (I believe) be restated in terms of functional equivalences.

III. EXTERNAL CONDITIONS

What, then, would it take for representation to occur? What sort of conditions need to be met for one agent, *A*, to act in the name of another, *B*? In this Section, I consider two “external conditions” that have been put forward in the literature. They are “external”

²¹ Alan Turing, “Computing Machinery and Intelligence,” *Mind* 59, no. 236 (1950): 433, <https://doi.org/10.1093/mind/LIX.236.433>.

²² Robillard, “No Such Thing as Killer Robots,” 707; Heather M. Roff, “Killing in War: Responsibility, Liability, and Lethal Autonomous Robots,” in *Routledge Handbook of Ethics and War: Just War Theory in the Twenty-First Century*, eds. Fritz Allhoff, Nicholas G. Evans and Adam Henschke (Abingdon: Routledge, 2013), 356. A different view comes from Luciano Floridi, “AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis,” *Philosophy & Technology* 38, no. 30 (2025): 1–27, <https://doi.org/10.1007/s13347-025-00858-9>, who argues that AI systems are full agents.

in the sense that they simply relate to the purported representatives' observable behavior, in contrast to their inner mental states.

At least one external condition seems necessary for representative agency. To see this, consider:

Rogue Mission: An elected politician, Arnold, goes on a rogue mission to assassinate a neighboring country's leader. When this is caught, he claims that he is acting on his democratic mandate.

Clearly, we would not want to say that Arnold is acting in the name of the people when he is on his mission. This is because there is no reasonable interpretation of his mandate which this action forms a part of. Even if assassinations of foreign leaders could have been part of a plausible interpretation of Arnold's electoral platform, it is only reasonable to expect that he would have pursued this policy through the institutions of government. We might thus, as a number of authors have explicitly done, place the following condition on representative agency:

1. The mandate condition: For A to act in the name of B, A must act within their mandate.²³

It seems perfectly possible for an AI system to meet the mandate condition. So long as the system only makes decisions regarding matters that we have authorized it to, the condition would be satisfied. This could be done, for example, through top-down programming rules that restrict the sorts of actions that the system can take.

Care needs to be taken, however, once we are dealing with systems that operate with higher levels of autonomy (particularly if they approach general AI, exhibiting human-level intelligence or higher across a wide range of domains). As Nick Bostrom illustrates, using a series of amusing examples, ensuring that these sorts of systems act within the bounds we intend will require careful specification of the task (which Bostrom ultimately seems to think might be impossible).²⁴ A superintelligent AI system given the innocuous-sounding task "put a smile on everyone's face," for instance, might accomplish this by performing involuntary plastic surgery on all humans to give

²³ Chiara Cordelli, *The Privatized State* (Princeton: Princeton University Press, 2020), 169; Arthur Ripstein, *Force and Freedom: Kant's Legal and Political Philosophy* (Cambridge, MA: Harvard University Press, 2009), 254.

²⁴ Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 120–122.

them a constant grin.²⁵ In one sense, of course, the system does act within the mandate that it has been given – it literally does what it has been programmed to do – but we might want to adjust the mandate condition to require that implicit understandings of the mandate are also followed. This is indeed what Chiara Cordelli does, arguing that representation may require acting “within the authorized domain of action, D, according to a *reasonable interpretation* of [the principal’s] own understanding of the boundaries of D at the time of the authorization.”²⁶ The mandate condition might thus be re-formulated as:

1*. The reasonable interpretation condition: For A to act in the name of B, A must act within a reasonable interpretation of their mandate.²⁷

Strictly speaking, Cordelli does not view the reasonable interpretation condition as a necessary condition. As we will see in Section V, she thinks that A might act in the name of B even if it is not satisfied, so long as some additional conditions are satisfied. For the moment, however, we can note that there will be at least some cases where an AI system acts within such a reasonable interpretation. Perhaps in some cases where a much greater degree of discretion is granted, this condition will be violated. But at least sometimes the condition will be met. When it is not, AI systems may well still be able to act in the name of humans, so long as Cordelli’s other conditions are met (and assuming that, when taken together, these are sufficient for representation).

Another possible necessary condition is put forward by Cordelli. She argues that there are some tasks which are of a nature such that any delegation to another agent is invalid. For example, she argues that we cannot validly abdicate our right to collective rule and, consequently, privatization of public services (beyond a certain threshold) involves invalid authorization of public decision-making to private actors.²⁸ To a version of the mandate condition, we might thus add:

2. The authorization condition: For A to act in the name of B, B must *validly* have delegated the authority to act.²⁹

Assuming that the decisions that we can validly authorize others to make is not an empty set, this condition poses no principled barrier to the possibility of an AI system

²⁵ Ibid., 120.

²⁶ Cordelli, *The Privatized State*, 169, emphasis added.

²⁷ Ibid.

²⁸ Ibid., 119–55.

²⁹ Ibid., 169.

acting in our name than the mandate conditions. So long as AI systems are used to make decisions that we have the moral power to delegate, the condition would be met. We simply need to limit their use to certain contexts.³⁰

In sum, there is no principled reason why an AI system cannot meet previously-identified external conditions on representative agency. This should be no surprise: since these conditions place limits only on the externally observable behavior of representative agents, it is difficult to see how at least some AI systems will not meet them just as well as human representatives. Of course, whether or not particular AI systems *will* meet the conditions, is another matter. But I conclude this Section by suggesting that there is no barrier posed by the external conditions for the possibility of AI acting in our name. As we will now see, however, externally-observable behavior may not be sufficient for genuine representation.

IV. EXTERNAL CONDITIONS

Why might the internal states of purported representatives affect whether or not they act in the name of others? Consider the following example from Cordelli:

Whistle-Blowing: British Petroleum (BP) hires a new spokesperson, Liz. Liz's mandate is to represent the company's official view on climate change. The official view is that BP is committed to a lower-carbon future. Liz is an environmentalist activist in disguise, and her aim is to blow the whistle by revealing BP's nefarious environmental practices to the public. However, because of a distraction, Liz ends up inadvertently communicating BP's official view to the public.³¹

According to Cordelli, despite acting within a reasonable interpretation of a validly-authorized mandate, Liz fails to speak in the name of BP. In cases like this, she claims, "the principal would be entitled to disavow their actions."³²

Why is this? Cordelli identifies the problem with the fact that agents like Liz comply with their mandates "either coincidentally or accidentally."³³ She consequently adds the following condition for representative agency:

³⁰ It may be claimed that the way in which we authorize others to make decisions may affect whether the authorization was valid. For instance, it may be thought that we can only validly authorize others if we retain power to guide the delegates or withdraw our authorization. For clarity, however, I will simply stipulate that the authorization condition rules out only authorizations based on what sort of decision is being delegated. These broader concerns are better subsumed under other conditions, to be discussed in subsequent sections.

³¹ Cordelli, *The Privatized State*, 161 (quoted verbatim).

³² Ibid.

³³ Ibid., 169.

3. The intention condition: For A to act in the name of B, A must carry out their mandate intentionally.³⁴

Liz fails to act in the name of BP because she does not intend to carry out her mandate.

Can an AI system meet the intention condition? To answer this question, we would need to consider the internal capacities of AI systems. More work in this area needs to be done, but one might reasonably doubt whether an AI system can even have intentions at all. At the very least, we might think it doubtful that an AI system can act for the right sort of intention, namely an intention to comply with the relevant mandate. It seems, rather, any intention would be cashed out in terms of the *content* of the mandate. The fact that something is a mandate would not appear to be relevant in AI decision-making.

While I think that Cordelli is correct to diagnose the problem with cases like *Whistle-Blowing* as the coincidental or accidental nature of the compliance, there is space between this and thinking that a mandate must be intentionally complied with for representation to occur. What seems crucial is that there is a sort of robust connection between the mandate and the compliance. In other words, it must be the case that compliance must not only be achieved in existing circumstances, but rather over a wide range of counterfactual circumstances.³⁵ In the case of *Whistle-Blowing*, this would include the counterfactual where Liz is paying due attention to what she is doing (in which case, she would not comply with the mandate). Robustness will generally be achieved when humans are representatives if they have the right sort of intention, but when other entities are in play there are alternative ways.

Perhaps, then, the more general condition that Cordelli is looking for is the following:

3*. The robustness condition: For A to act in the name of B, A must be robustly disposed to carry out their mandate.

An AI system can meet this condition. All that is required is an insensitivity to various variables changing for acting within the mandate. We might describe this insensitivity as a functional equivalent (in the sense outlined earlier) of an intention to comply. One might be worried at talk of A being “disposed” to do something here. Can AI systems even have dispositions? But this should not be understood in too literal a way; it simply refers to a reliable connection between the mandate being given and the mandate being carried out.

³⁴ Ibid.

³⁵ On the value of robustly-demanding protections more generally, see Philip Pettit, *The Robust Demands of the Good: Ethics with Attachment, Virtue, and Respect* (Oxford: Oxford University Press, 2015).

Why do I favor the robustness condition over the intention condition? I would suggest that what is doing the intuitive work in *Whistle-Blowing* is the worry that, by having a spokesperson like Liz, we come worryingly close to being on the hook for all sorts of actions and speech acts to which we have no special connection. What representation is supposed to do is, to return to Pitkin's insight, "make the represented present again" in a sphere of delegated action. What this requires is a certain reliability that one's wishes will be carried out. Intention seems superfluous to this purpose.

To motivate why some think that a further internal condition is required, recall that many purported representatives must be given a degree of discretion in what they do within their mandate. But there may be illegitimate ways of exercising that discretion if one's representative status is to be maintained. Consider the following case, again from Cordelli:

Infiltrating Spy: A KGB spy infiltrates the US government as a public official. In order to avoid being discovered, the spy must do his or her job as a US official impeccably. Therefore, the spy intentionally aims to do everything the spy's mandate as a public official requires him or her to do.³⁶

What might seem to have gone wrong here is that the spy does not act for the right sort of *reasons*. We might think that, to be a genuine representative, one must act on reasons that are shared with those represented. If we are attracted to this view, we would want to add the following condition, entertained by Cordelli:

4. The shared reasons condition: For A to act in the name of B, A must act on the basis of reasons that are shared by B.³⁷

The reason why the spy is not a genuine representative is that the US public (whom they purport to represent) do not share the reasons for acting, namely, the advancement of the goals of the Russian security services.

It is questionable whether an AI system could meet the shared reasons condition. For one thing, it is questionable whether existing and near-future AI systems can act on the basis of reasons at all. Even if they can, however, the sorts of reasons for which they act would most likely be very different from those that humans would have. Artificial neural networks, for example, operate by breaking up inputs into large numbers of variables, and combining these variables in a function that gives an output

³⁶ Cordelli, *The Privatized State*, 162, (quoted verbatim).

³⁷ *Ibid.*, 163.

value. While such networks operate in similar ways to the workings of the human brain, there are nonetheless important differences. A human, for example, when asked why they fired their weapon might say “because there was a tank ahead.” An autonomous weapon system, on the other hand, is likely to lack the concept of a tank in all but a metaphorical sense; explanations for why it fired would most likely take the form of pointing to various patterns that were detected (which are correlated with enemy tanks being ahead).³⁸

How might we respond to this potential challenge to the possibility of AI representatives? Cordelli, after introducing it, has argued that the shared reasons condition is too strong; all that is required is the absence of “excluded reasons” (such as wanting to help a rival government).³⁹ An AI system would presumably be able to meet this condition. (If it lacks all reasons, it will certainly meet it.) But I think that we can argue for a less radical revision that does justice to the intuitive thinking behind it.

We can return to Pitkin’s point, mentioned above, that the purpose of representation is to make the represented present again. Cordelli argues that it is this insight that might lead us to adopt the shared reasons condition.⁴⁰ It is easy to see why this might lead one to adopt the shared reasons condition: representatives who act on shared reasons act in ways that the represented would act if they were present. But there are other ways in which we can achieve this without relying on shared reasons.

To see this, we can begin by making the distinction between causal reasons and motivating reasons. Causal reasons are simply the causes of certain actions and events: a cause of a forest fire might be a bolt of lightning hitting a tree. Motivating reasons are a class of causal reasons: those that relate to intentional agents’ mental states. The motivating reason of my putting the forest fire out is my desire to save my log cabin. On this account, all motivating reasons are causal reasons, but not all causal reasons are motivating reasons.⁴¹

³⁸ Herman Cappelen and Josh Dever, *Making AI Intelligible: Philosophical Foundations* (Oxford: Oxford University Press, 2021), 81–102, have argued that we can attribute content to the outputs of AI systems and, in particular, we can view predicates in their outputs as meaning what we mean if they were trained on a data set that was hand-coded using the predicate. Whether or not the internal representations of AI systems can also be said to contain predicates like “is a tank” is another question, however, and one that has, to my knowledge, not been explored in any great detail. I will assume that this is not the case here, and explore if AI systems can still represent us even if they lack such concepts in their internal representations.

³⁹ Cordelli, *The Privatized State*, 166.

⁴⁰ *Ibid.*, 163.

⁴¹ Cf. Maria Alvarez and Jonathan Way, “Reasons for Action: Justification, Motivation, Explanation,” *Stanford Encyclopedia of Philosophy*, 2024, <https://plato.stanford.edu/entries/reasons-just-vs-expl/>.

The shared reasons condition seems to suggest that the *motivating* reasons for a representative acting must be such that they could be shared by those represented (these reasons would cohere with to their value system, contingent ends, etc.). The problem that AI brings up is that it is unclear whether artificial systems can act have motivating reasons (or at least the same motivating reasons). But the outcomes of AI decision-making certainly have causal reasons: these include the inputs into the system and the way in which they are combined in algorithmic decision-making. My suggestion here is that representation can occur if these causal reasons are suitably related to the motivating reasons that the represented can share.

Here, more precisely, is my proposal for an alternative condition to the shared reasons condition:

4*. The supervenient reasons condition: For A to act in the name of B, A's actions must have causal reasons that supervene on motivating reasons that are shared by B.

When I talk about a causal reason supervening on a motivating reason, I mean that there can be no change in the causal reason without a corresponding change in the motivating reason. Thus, for this condition to be met, the causes of A's actions changing will always be accompanied by a change in B's motivating reasons.

There are two things to note about this condition. The first is that it captures the intuitive grounding of the shared reasons condition just as well as that condition. If the goal of representation is to make the represented present again, there is a sense in which the represented are present if our motivating reasons are sufficiently connected to the causal reasons why the representative acts. For example, if we set up a system of government that uses incentives (financial and reputational) to induce government officials to act in our interests, their causal reasons (which are also motivating reasons) for action may be sufficiently related to our own preferences and values to make it appropriate to say that we are "present" in government. The way we set up the incentive structure is determined by what we want the officials to do. If we wanted them to act in different ways, we would set up the incentives differently. At the very least, it seems, we are represented here, so long as the incentive structure is well designed.

The second thing to note is that it can, in principle, be met by an AI system. If my motivating reason for wanting artillery to be fired in a battlefield is if there is an enemy tank ahead, and the causal reasons why an AWS fires is because of the way I set up its programming, the condition is met. If there were a change in my motivating reasons, I would set up the system differently, and the causal reasons for action would change. If this condition is right, it suggests that such an AWS can act in my name, even if it lacks

motivating reasons (or lacks the same motivating reasons, which might be the case if it lacks the concept of a “tank,” for instance).

It may be objected that the condition is too strong for AI systems to meet. After all, autonomous systems, as we have seen, often act in unpredictable ways not fully programmed in advance. How are we to ensure that they have the right sort of causal reasons? This would be a significant engineering challenge, but I do not think that these considerations tell against the *possibility* that AI systems could meet the condition. It is plausible to think that an autonomous system might have programming such that its causal reasons are of the right sort. We might accomplish this by testing the system over a wide range of training scenarios, and adjusting the system when its behavior deviates from what we would want. If the training scenarios are sufficiently varied, we might ensure that there is a high probability that the system would act in our name on each occasion that it makes a decision. How often this will in fact occur is not something that I can speculate on here. The purpose of the paper is more modest: to establish the possibility of representative robots. These considerations suggest, however, that the number of occasions in which an AI system acts in the name of others may be not uncommon, so long as an adequate testing regime is in place.

A related worry, however, is that we could never *know* whether the condition is met. Despite exhibiting behavior that looks like it is met in training, the system could be quite different than we assume. The same is true, of course, of human representatives: we have even less influence over the sorts of reasons they act on (we do not program them in advance, as we do AI systems). And we may never know their true reasons for acting. For all we know, some of those in our government might be acting on the wrong sort of reasons, as Cordelli’s spy does. This is therefore not a unique obstacle for AI representation, but for representation more generally, assuming our theory of representation has some condition relating to the reasons that a representative acts. The most we can hope for in practice, I think, is a high degree of confidence that the condition is met. If that is the case, we are justified in acting as if the system represents us.

One complication that might arise here, especially if we are dealing with AI systems in government functions, is how we are supposed to attribute motivating reasons to the represented party, since in this sort of case this is likely to be a collective of individuals, each of whom may have different reasons. Those developing theories of representative agency sometimes argue that there needs to be a way of determining the “general interest”⁴² or “omnilateral will”⁴³ that is to guide representatives. This may

⁴² Dorfman and Harel, *Reclaiming the Public*, 107.

⁴³ Cordelli, *The Privatized State*, 17.

seem impossibly utopian, in which case the idea that AI systems can act in the name of groups of people might be questioned (but so would the more familiar idea that agents of the state can act in the name of the people). Even if this is the case, however, the idea that AI systems can act in the name of *individuals* would still be possible. This result is not to be underestimated.

Overall, then, while a number of possible internal conditions on representative agency might not be satisfiable by AI systems, I have suggested that we can legitimately generalize or weaken these conditions in a way that makes it possible for AI systems to meet them.

V. RELATIONAL CONDITIONS

Is some combination of external and internal conditions necessary and sufficient for proper representation to exist? Cordelli argues not. She provides the following case to illustrate why:

Web Designing: I hire a web designer, Tom, to build my academic website...Tom has worked for Hollywood clients for many years. His understanding of what counts as a successful website is strongly affected by his previous career experience and his previous clients' tastes. Upon reflection, he decides to post a picture of an attractive actress next to each of my academic papers, so as to solicit as many views as possible. Nothing in the contract explicitly prohibited him from doing so.⁴⁴

Tom's actions might well meet a combination of some of the previous conditions we have discussed. He acts within his mandate. Delegating the task of designing a website seems to be something that someone is quite entitled to do. It is quite plausible to suppose that Tom both intends to carry out the mandate and does so for reasons that are not excluded. More generally, he is robustly disposed to carry out the mandate, and his reasons for doing so may supervene on his employer's motivating reasons. Yet we may still want to say that Tom did not speak or act for his client.

At least one other condition is needed to explain this result. As we have seen, Cordelli argues that a sufficient condition might be that the representative act on a reasonable interpretation of the mandate (what I called the reasonable interpretation condition). Since Tom does not act on a reasonable interpretation of the mandate, he is not a genuine representative. But Cordelli argues that this is only one of three additional conditions that might be added to her preferred combination of external and internal conditions.

⁴⁴ Ibid., 167 (quoted verbatim).

How else might Tom gain representative status? Cordelli argues that what would be needed, is “some more active involvement on my part.”⁴⁵ If the delegate has the possibility of approving a decision before it becomes consequential (e.g. reviewing a designed website before it went live), there would be genuine representation here. Nonetheless, this might not always be possible: perhaps the delegate must make consequential decisions before any review process has taken place. “In the case of decisions that are not easily or quickly verifiable,” says Cordelli, “we need a mechanism of *ex ante* consultation, beyond a method of *ex post* review.”⁴⁶ Either of these possibilities – requiring *ex post* review and requiring *ex ante* consultation – might be the sort of additional condition we are looking for. The idea here, as I understand it, is that a discussion between a representative and those they represent prior to acting can help the latter to clarify the broad principles on which they would like to guide the former’s action can help to clarify ambiguities in the mandate.

If we are dealing with autonomous AI systems, *ex post* review may often be impossible. As we have seen, when systems are autonomous in the second sense of the term I introduced in Section II, it may be infeasible to allow humans to review decisions before they are actioned, owing to time and other constraints. Moreover, given the counter-intuitive ways in which AI systems often interpret their instructions (recall Bostrom’s example of the robot who performs involuntary plastic surgery on people), we cannot be sure that they will always act on a reasonable interpretation of their mandate. It may be most promising, then, if we want representative AI systems, to focus on fulfilling the condition relating to consultation.

Other theorists also emphasize the need for consultation. They differ, however, in their views on what the consultation has to be like. Take, for instance, the theory of Wendy Salkin in her study of informal political representation. Salkin argues that one way in which a group might authorize an informal political representative, that is, a representative which does not operate in standard political institutions like parliaments, is through, first, achieving “provisional authorization” and, second, strengthening their normative power through a process of consultation. Someone possessing provisional authorization

does not yet have the group’s trust or wholehearted support. So, before they go out and speak or act on the group’s behalf (for instance, calling a press conference or issuing a public statement), they must make reasonable efforts to consult

⁴⁵ Ibid., 168.

⁴⁶ Ibid.

known and accessible group members to make explicit both their plan and its justification.⁴⁷

The normative powers that such a representative would have might then increase if they receive support from those they represent in this consultation.⁴⁸

Avihay Dorfman and Alon Harel's theory also has a central place for consultation. For them, what distinguishes a genuine (political) representative whose acts constitute acts *of the state* from someone who merely acts *for the state* is that the representative operates with a "deferential" concept of fidelity.⁴⁹ What this means is that, to act in the state's name, "an assistant must suppress his or her own judgment...and accept [the state's] judgment."⁵⁰ The fact that the state demands something should be a pre-emptive reason,⁵¹ whereby all other reasons for action (including, for example, the inherent desirability of different courses of action) lose their status as reasons because the state demands action of some sort.

In practice, Dorfman and Harel argue, ensuring that political representatives act according to a deferential conception of authority will require creating a "community of practice" of which the representatives form a part. This, in turn, requires two things. First, it requires the existence of an "institutional structure in which the general interest – as seen from the public point of view – is articulated."⁵² And, second, it requires an "integrative practice" where the public officials who claim to speak for the public are integrated into the political community whose interest is the public one.⁵³ Crucially, this second aspect involves the possibility of those whose name is acted in to "change the practice, guide its mode of operation, and reevaluate the norms governing it."⁵⁴ This, too, is likely to involve some consultation.

What all this suggests, then, is that one important way in which purported representatives can ensure that they are acting in the name of individuals is through consultation with them. Given the discretion that representatives inevitably have, detailed instructions covering every eventuality cannot be given in advance (if such instructions could be given, we would probably be talking about proxy agency rather

⁴⁷ Wendy Salkin, *Speaking for Others: The Ethics of Informal Political Representation* (Cambridge, MA: Harvard University Press, 2024), 117.

⁴⁸ Ibid.

⁴⁹ Dorfman and Harel, *Reclaiming the Public*, 100.

⁵⁰ Ibid.

⁵¹ Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1986), 57–62.

⁵² Dorfman and Harel, *Reclaiming the Public*, 107.

⁵³ Ibid., 108.

⁵⁴ Ibid., 109.

than representation). Instead, discussions need to arrive at broad principles on which the representative is going to act. For Cordelli, this simply needs to be taken before actions are taken. The theories of Salkin and Dorfman and Harel, on the other hand, seem to require a more demanding back and forth on a more regular basis.⁵⁵

This suggests a fifth condition for representation to occur:

5. The consultation condition: For *A* to act in the name of *B*, *A* must consult with *B* about the principles guiding their action, and be open to adjusting their behavior in light of this.

To be clear, the consultation condition may not be a necessary condition.⁵⁶ But, given how autonomous systems are likely to make decisions, it may be the only sufficient condition that is possible to meet in a range of cases.

But what would consultation look like exactly? With familiar forms of representation by humans, it might require forms of deliberation between the represented and the representative. The latter could make explicit the guiding principles of their action through discussing real-world cases that they have knowledge of. The represented, in turn, could evaluate these principles and, where they see fit, call for different principles.

Can something similar occur when the representatives are machines? To be sure, the very idea of deliberation with a machine may strike us as a non-starter. But, I want to suggest, we might achieve the same results by allowing the examination, evaluation, and adjustment of the algorithms that underlie them. Is this feasible? In their discussion of this point, Dorfman and Harel raise doubts about this. They argue that the opacity of these systems renders any decisions made by them non-transparent, unpredictable, and unchallengeable.⁵⁷ These are thus not the sorts of decisions that we can guide, given our lack of knowledge.

In what follows, I want to suggest that Dorfman and Harel's conclusion is premature. Crucially, they have overlooked the potential uses of explainable AI (XAI) in ensuring a system of evaluation of AI decision-making. XAI refers to a cluster of techniques that, in one way or another, make opaque AI systems understandable to humans. They might do this, for example, by providing approximations of how the systems work, by furnishing

⁵⁵ Cordelli, *The Privatized State*, 159–60 argues that this is too strong: it seems to be a necessary condition for proxy agency rather than representative agency.

⁵⁶ As we have seen, Cordelli argues that alternative conditions could be met. And Salkin also suggests that something much weaker, such as “tacit authorization” (i.e. the absence of objections), might even be enough. Salkin, *Speaking for Others*, 116–17. Recall also from earlier there are different types of representation: consultation might only be important for some of these.

⁵⁷ Dorfman and Harel, *Reclaiming the Public*, 171–87.

examples of how they worked in concrete cases, or by specifying counterfactuals, that is, by explaining how adjusting one of the input variables would affect the output variable.⁵⁸ These techniques may be used in combination. My suggestion is that, by making use of one or more of these techniques, we can become sufficiently informed about autonomous AI systems in order to conduct a meaningful evaluation and, if necessary, recommend an adjustment of the system in question. At least, we can be hopeful that this will help in some contexts.

To take a very simple example, suppose that, upon purchasing a self-driving car, we are interested in the sorts of principles it works with in accident scenarios, such as how will it behave in the face of an accident when at least one person will be harmed. Suppose that we are provided with a previous example, where the system prioritized the life of a lone passenger over two pedestrians. Pressing the manufacturer further about why the system behaved in this way, we are provided with counterfactual reasoning: the two pedestrians were designated as jay-walkers (they were crossing the road on a red light), yet if this designation was not made, the car would prioritize their lives. It becomes clear to us that the system is operating in line with a principle stating that, when some individual is morally responsible for a threat of harm occurring, they are morally liable for that harm being redirected to them.

Such a principle, if we could program it into a self-driving car, might strike us as reasonable.⁵⁹ But suppose that you do not endorse this principle. Perhaps you would prefer simply that, in accident scenarios, we minimize the overall harm to humans (even if this means that you will be at greater risk of being harmed). If the manufacturer offers to reprogram the car more in line with your values, it seems to me, the car might be said to act in your name. The relational condition is met here.

XAI, then, looks like a promising tool for the consultation condition to be met. So long as humans are given the right sort of information, and the opportunities for adjusting the system, something serving the same function as deliberation in other cases of representation can take place.

Are there any other relational conditions? Salkin argues that, for a (formal) political representative to be authorized to speak in the name of a constituency, “the constituency will at some future point have an opportunity to hold their FPR [formal political representative] accountable for the FPR’s decisions by means of a different

⁵⁸ Timo Speith, “A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2022).

⁵⁹ For something like this view, see Antti Kauppinen, “Who Should Bear the Risk When Self-Driving Vehicles Crash?” *Journal of Applied Philosophy* 38, no. 4 (2020): 630–45, <https://doi.org/10.1111/ja12490>.

procedure [than deliberation], like a subsequent election.”⁶⁰ This suggests a final condition on one agent acting in the name of another:

6. The recall condition: For A to act in the name of B, B must have an opportunity to stop A from being recognized as acting in their name.

Once again, this condition imposes no principled reason why an AI system cannot act in the name of others. So long as humans can choose, at certain points, to stop the system from operating, the condition will be met. Of course, they may not be able to do this at all times. Military commanders may need to wait until an AWS returns from battle before shutting it down for good, owing to the inability to communicate in battlefield settings, as discussed above. The requirement of having a “human-in-the-loop,” having to approve of every proposed action by an AI system in advance, is often an unobtainable ideal.⁶¹ But this is no less true of citizens and their (human) political representatives. Opportunities for recall may only arise every few years when an election comes around.

Perhaps, however, the motivation for adding a condition like this is that, when it is met, purported representatives have incentives to follow the wishes of those they seek to represent. If politicians do not do this, for instance, they may be voted out of office. It may look unlikely that an AI system would respond to incentives in the same way (no existing or near future AI system is likely to reason in such a sophisticated manner about ways to prevent themselves from being shut down). Consequently, it may be thought, even if the condition is technically met in a case involving an AI system, it is not met in a way that contributes towards the system’s representative status.

Two things can be said in response to this objection. First, if this is the thinking behind the recall condition, then I am no longer sure that it is a genuine condition on acting in the name of others. It rather looks like a useful way in which we can ensure another condition – the supervenient reasons condition – can be met in some cases. And, while it might be virtually indispensable when we are dealing with human representatives, I suggested in Section IV that this condition can be met by AI systems

⁶⁰ Salkin, *Speaking for Others*, 114. Note that, for Salkin, an agent can gain representative status merely by being recognized by an audience as having it. She then specifies additional conditions that need to be met for an agent to be an *authorized* representative, with normative powers. I have treated these conditions as conditions to be a representative in the first place in this paper, but readers sympathetic with Salkin’s framework can think of these as conditions for authorized representation.

⁶¹ Sparrow, “Killer Robots,” 68–69.

when they are programmed in a way that is sufficiently sensitive to certain reasons that those represented have. Providing a system of carrots and sticks is not needed.

Second, even if this condition and this rationale are correct, something very much like rewards and punishments can be imposed on AI systems. Turing uses precisely the language of punishments and rewards when discussing how machines might learn.⁶² The idea would be that machines can be sent “reward signals” and “punishment signals,” and their programming would lead them to become more likely to repeat the decisions that preceded the former and less likely to repeat the decisions that preceded the latter. Once again, to return to one of my earlier points, we should not let our anthropocentric concepts and theories of representation, which were developed with human representatives in mind, prejudice the question of whether AI systems could act in our name.

VI. CONCLUSION

The possibility that AI systems can act in our name, I have claimed, offers a novel and promising solution to various worries about their use in high-stakes decision-making. I have offered a provisional assessment of their ability to do so by considering different conditions, combinations of which have been thought necessary in order for one agent to represent another. While I have suggested that there are grounds for thinking that AI systems can indeed represent us, either because they can meet the conditions or because the conditions need to be revised in light of new possibilities opened up by AI (quasi-)agency, further work is needed in a number of directions before we arrive at the comforting conclusion that building representative robots is a silver bullet capable of assuaging a number of different worries about the deployment of AI.

First, what conditions need to be met for one agent to act in the name of another more generally is an underexplored question. Perhaps there are plausible conditions not discussed here. If that is the case then, depending on the nature of these conditions, we might in turn need further exploration of how AI works in order to see if they can meet these conditions. If they are internal conditions, for example, we might need to consider what internal mental states (such as beliefs, desires, and intentions) AI systems are capable of – this was a question that I was largely able to side-step in this paper.

Second, while this paper has argued that AI systems *could* act in our name, how often the conditions for this can be met is another question. Further investigation, for instance, into what sort of information XAI techniques can provide us – and, in

⁶² Turing, “Computing Machinery and Intelligence,” 457.

particular, whether considering and adjusting AI systems based on the sort of knowledge and understanding that XAI can bring – is needed.

Despite these qualifications, the general strategy that this paper points towards – trying to build representative AI systems rather than AI systems that we can directly control – may be a fruitful avenue to explore. This is true especially in light of the apparent failures of the “control strategy” to ensure systems that protect responsibility, ensure meaningful participation, and promote human values.

At the same time, the normative consequences of the deployment of representative AI systems might look very different than if control were exercised. If there are AI systems that are controlled by users or developers, for instance, responsibility for their behavior would lie with this (relatively small) group. If AI systems act in the name of others, however, responsibility would be shared among all the individuals who make up the represented parties (which might involve all members of a country). Viewing AI systems as representatives rather than tools in our control may thus have significant upshots with normative significance.

Acknowledgements

I would like to thank audiences at the MINT-Yale Workshop on the Normative Philosophy of Computing and the Department of Philosophy at the University of Vienna, as well as the editors and reviewers of this journal, for feedback on previous drafts.

Competing Interests

The author has no competing interests to declare.

