AI & SOCIETY
https://doi.org/10.1007/s00146-023-01698-x

OPEN FORUM

Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers

Robert Sparrow[1]

Received: 11 September 2022 / Accepted: 22 May 2023

Abstract
When asked about humanity's future relationship with computers, Marvin Minsky famously replied "If we're lucky, they might decide to keep us as pets". A number of eminent authorities continue to argue that there is a real danger that "super-intelligent" machines will enslave—perhaps even destroy—humanity. One might think that it would swiftly follow that we should abandon the pursuit of AI. Instead, most of those who purport to be concerned about the existential threat posed by AI default to worrying about what they call the "Friendly AI problem". Roughly speaking this is the question of how we might ensure that the AI that will develop from the first AI that we create will remain sympathetic to humanity and continue to serve, or at least take account of, our interests. In this paper I draw on the "neo-republican" philosophy of Philip Pettit to argue that solving the Friendly AI problem would not change the fact that the advent of super-intelligent AI would be disastrous for humanity by virtue of rendering us the slaves of machines. A key insight of the republican tradition is that freedom requires equality of a certain sort, which is clearly lacking between pets and their owners. Benevolence is not enough. As long as AI has the power to interfere in humanity's choices, and the capacity to do so without reference to our interests, then it will dominate us and thereby render us unfree. The pets of kind owners are still pets, which is not a status which humanity should embrace. If we really think that there is a risk that research on AI will lead to the emergence of a superintelligence, then we need to think again about the wisdom of researching AI at all.

Keywords Artificial intelligence · Ethics · Superintelligence · Republicanism · Bostrom · The singularity

When asked about humanity's future relationship with computers in 1970, Marvin Minsky, one of the founding fathers of the field of artificial intelligence, famously replied, "Once the computers got control, we might never get it back. We would survive at their sufferance. If we're lucky, they might decide to keep us as pets" (Darrach 1970, 68). This prospect has led a significant proportion of those concerned with the "existential risk" posed by AI to conclude that there is an urgent need to confront what has become known as the "Friendly AI" problem. How can we make sure that future "superintelligent" computers *do* care about our welfare? In this paper, I draw on the political philosophy of republicanism to argue that this project misses a fundamental—indeed in some respects, an obvious—point about the future that Minsky is imagining, which makes it profoundly troubling. A key insight of the republican tradition is that freedom requires equality of a certain sort, which is clearly lacking between pets and their owners. We should try to avoid being in a position where we need to be "lucky" to survive not just because our luck might run out but because to depend on the sufferance of superiors is to be enslaved. For this reason, solving the Friendly AI problem would not make the prospect of becoming AI's pet any more attractive. To the extent that we think there is a real risk that machines will become superintelligent, a concern for human freedom gives us strong reason to halt research that might lead to the emergence of superintelligent AI.

1 Superintelligence and the Friendly AI problem

While the view is not shared universally amongst computer scientists and engineers, a significant number of eminent authorities, with doctorates in computer science, physics,

✉ Robert Sparrow
robert.sparrow@monash.edu

[1] Department of Philosophy, Faculty of Arts, Monash University, Clayton, VIC 3800, Australia

Published online: 13 June 2023

mathematics, and philosophy, believe that there is a real danger that artificial intelligences will enslave—perhaps even destroy—humanity. For instance, Stuart Russell, co-author of a leading textbook on AI is, in his own words, "publicly committed to the view that [his] own field of research pose[s] a potential risk to [his] own species" (2019, 4). Max Tegmark, a physicist based at MIT and one of the founders of the Future of Life Institute, suggests, in the course of talking about the implications of AI, that "it could be that machines themselves outsmart us and manage to take control" (Anthony 2017). Nobel prize-winning psychologist Daniel Kahneman, in an interview in the Guardian, observed that "clearly AI is going to win [against human intelligence]. It's not even close" (Adams 2021). Ray Kurzweil, a talented engineer who did pioneering work in optical character recognition, has had a profitable second career warning humanity about the coming "Singularity" in which humanity must cede its place to superior machines (Kurzweil 2000, 2005).

Oxford University philosopher, Nick Bostrom, has written an entire book, *Superintelligence*, discussing the threat posed to humanity by artificial intelligence (Bostrom 2014). Bostrom defines a superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2014, 22). An important reason why, Bostrom believes, we should be worried that research on AI will lead to the emergence of superintelligent AI is that the creation of AI involves the risk of an "intelligence explosion" (Good 1966). If we, with human level intelligence, are capable of creating an AI that is, or can become, slightly more intelligent than us, then that AI may be able to create an intelligence that is slightly more intelligent than it, which in turn may be able to do the same… and so on, until machines are superintelligent (Chalmers 2010). Because the intelligence of AI is likely to be a function of its software, this process of iterative improvement may occur very quickly, perhaps even before human beings have the opportunity to interrupt it (Yudkowsky 2008).

According to Bostrom, a super-intelligent AI might cause the extinction of humanity, or at least a large portion thereof, either because its goals evolve in such a way that it comes to see us as a threat, or at least a nuisance, or because we fail to anticipate how the goals that we grant AI might lead to it acting in ways that result in our extinction. This latter possibility is the motivation behind Bostrom's notorious "paper clip maximiser" thought experiment, wherein we are asked to imagine the possibility that an AI we built to manufacture paperclips would use us as raw materials for this process without realising that this choice would vitiate the reasons why we asked it to make paperclips in the first place (Bostrom 2014, 123–125).

One might think that it would swiftly follow that we should abandon the pursuit of AI. We should strive to bring it about that working to develop AI is akin to working to develop chemical or biological weapons, universally condemned, and prohibited under international law. However, this is not the conclusion reached by most people writing in this literature. Many seem to believe that anything that can be invented will be invented—and therefore that there is no point in trying to prevent the development of AI. A few seem to think that the benefits of AI justify the risks—although given that the risks involve extinction for humanity, those benefits would have to be substantial indeed.

Instead, most of those who purport to be concerned about the existential threat posed by AI default to worrying about what they call the "Friendly AI problem" or, occasionally, the problem of ensuring AI "value alignment" (Bostrom 2014; Gabriel 2020; Russell 2019; Yudkowsky 2001). Roughly speaking, this is the question of how we might ensure that what Hans Moravec (1988) describes as our "mind children"—the AI's that will develop from the first AI that we create—will remain sympathetic to humanity and continue to serve, or at least take account of, our interests (Russell 2019; Yudkowsky 2008). In order to dramatize what is at stake, I like to think of this as the "how do we ensure that AI does not eat us" problem.

It is difficult to know how seriously to take the debate about the existential risk posed by superintelligence. While one group of experts, which includes the figures discussed above, appear to be genuinely worried about "superintelligence", many computer scientists, and other people working in artificial intelligence, seem to think that the whole discussion is silly and that there is little prospect of machines becoming genuinely intelligent, let alone super-intelligent, for the foreseeable future.[1]

For the purpose of the current discussion, I am going to assume that there is a genuine prospect that "Strong" AI may be developed in the not-too-distant future and that there is a risk of an intelligence explosion leading to the evolution of superintelligent AI. That is to say, that there is a real danger that the situation that Minsky imagines will come about and that "if we're lucky" superintelligent machines will "keep us as pets". I want to argue that solving the Friendly AI problem would not make the prospect of becoming AI's pet any

[1] For a recent sceptical take, see Larson (2021). *IEEE Spectrum* surveyed the range of views on the prospects for superintelligence in a Special Issue on The Singularity in June 2008 (Vol. 45, no. 6). There is another paper to be written about why some technological possibilities are chosen to be the topic of public handwringing and others not—and when. No ink has been spilled about the danger that we will be displaced as a species by genetically modified super-intelligent tapirs, despite the fact that the creation of such creatures doesn't look any more impossible than the creation of super-intelligent AI. If everything that can be invented will be invented, one presumes we should be just as worried about this—and many other outré possibilities—as we are about humanity being made extinct by super-intelligent computers.

more attractive. A focus on the motivations of AI neglects the power that AI would have over us, the problematic status of pets, and the implications of both for our freedom.

2 A republican theory of freedom

To understand precisely *why* "AI's pet" is not a status to which humanity should aspire, it will prove useful to take a brief excursion into the "neo-republican" philosophy of Philip Pettit (1997, 2001, and 2012). Republicanism is well suited to the examination of ethical issues around AI because it is centrally concerned with the relationship between power, liberty, reason, and status. Pettit is the contemporary philosopher who has done the most to explain what republicanism says about this relationship.

The core intuition of republicanism, as represented by Pettit, concerns the nature of liberty. To be free, according to republicans, it is not enough that no-one prevents you from acting as you choose: one's freedom of action must be "resilient" or "robust". In particular, one is not free if one can only act as one wishes at the sufferance of the powerful. To be free, one must not be "dominated".

According to Pettit,

> "…someone has dominating power over another, someone dominates or subjugates another, to the extent that

1. They have the capacity to interfere
2. On an arbitrary basis
3. In certain choices that the other is in a position to make" (Pettit 1997, 52).

A key feature of this account of the relationship between power and liberty is that not all interference will count as dominating—the interference must also be *arbitrary*. Pettit suggests that,

> "An act is perpetrated on an arbitrary basis… if it is subject just to the *arbitrium*, the decision or judgement, of the agent; the agent was in a position to choose it or not to choose it, at their pleasure … in particular, since interference with others is involved, we imply that it is chosen or rejected without reference to the interests, or the opinions, of those affected. The choice is not forced to track what the interests of those others require according to their own judgements" (Pettit 1997, 55).

The recognition that the exercise of power is not arbitrary where it is required to track the interests of those affected allows republicanism to reconcile the laws of a well-ordered—that is, roughly speaking, a democratic—republic with the liberty of citizens.

Already, then, with just the republican account of the relationship between power and liberty in hand, we can see clearly what would be wrong with becoming AI's pet. The pets of kind owners are still pets and vulnerable to the whims of their masters, even if their masters treat them well. What is wrong with being someone's pet is not that they necessarily treat you badly, but that you are ultimately their toy and "free" to run and play only at their pleasure. Even if a pet's owner grants the pet the run of the house, to be a pet is to be enslaved.

3 Friendliness and freedom

It might be objected that a Friendly AI would only interfere in our lives in our interests and thus its exercise of power would not count as domination.

However, a benevolent dictator, who only interferes in our lives when it is in our interests, is still a dictator and his/her power is still inimical to our freedom. To be "free" to act as one wishes only at the sufferance of another is no freedom at all. As noted above, freedom requires that one's ability to act as one wishes is "resilient"—that it does not disappear immediately a more powerful party decides that it should. Thus, the exercise of power in accordance with our interests is only compatible with our freedom if we could resist it.

Pettit again:

> "What might make it possible for… a decision not to have the aspect of an arbitrary act of interference? The answer which suggests itself is: the fact that we can more or less effectively contest the decision, if we find that it does not answer to our relevant interests or relevant ideas" (Pettit 1997, 185).

Insofar as its purpose is to rule out the possession or exercise of power that does not track the interests of those it affects—which in the case of government power, is potentially anyone—this contestation must be:

> "…by recourse to public discussion in which people may speak for themselves and for the groups to which they belong. Every interest and every idea that guides the action of a state must be open to challenge from every corner of the society…" (Pettit 1997, 56).

According to republicanism, the exercise of power is—and is only—compatible with liberty when it is hostage to reason.

Perhaps, then, a genuinely Friendly AI would only interfere in the affairs of humanity after asking us what we want and listening to our deliberations? AI might serve as executive rather than legislature, efficiently bringing about what we want rather than deciding what would be good for us (Russell 2019).

There is something in this. Clearly, in many circumstances, AI can be used as a tool to help us realise our own goals. One way in which a super-intelligent AI might help us do this is by using its "intelligence" to determine the most efficient means to achieve our ends (Russell 2019).

However, ultimately, this argument founders—as does much of the discussion of Friendly AI—on the question of the relationship between the intelligence of AI, its power, its freedom, and *our* freedom. Both parts of the formulation "hostage to reason" in the republican account of how power must be limited in order to be compatible with liberty are important. In order for the power of the state to be compatible with the liberty of citizens, the government must listen to reason and justify its exercise of power in terms that citizens accept. However, it must also be hostage to reason in the sense that if the citizenry is not convinced that the government's exercise of power is justified, the government's power is checked. The ultimate check on the power of governments is the capacity of the citizenry to overthrow them.

Unfortunately, it is an article of faith in the literature on superintelligence that a super-intelligent AI would be able to use its superior intelligence to thwart human attempts to limit its power (Bostrom 2014, 91–104; Yudkowsky 2008). The Friendly AI problem arises precisely as a result of recognizing the power that the "super" intelligence of future AI will grant it. The only thing that will prevent a Friendly AI from eating us is if it does not want to. We would, therefore, remain subject to its whims and thus dominated and so unfree.

Granted, some formulations of the Friendly AI problem imply that solving it requires *guaranteeing* that AI will never act against humanity's interests (Russell 2019; Yudkowsky 2008). That is, in order to be truly friendly, it must be the case that a Friendly AI would never become unfriendly. Given the profound difficulties involved in constraining the activities of a superintelligence, this would require that it comes into existence with the desire to serve humanity's interests *and* that it never desires to change its own motivations (Russell 2019; Yudkowsky 2008).

The idea that we might be ruled by, or even live alongside of, a supremely powerful intelligence that is guaranteed not to act against humanity's wishes by virtue of never wanting to do so pushes republican intuitions about liberty, as well as our understanding of free will and possibility, to their limits. Nevertheless, it remains the case that a guarantee that AI *will not* "eat" us is not sufficient to establish that co-existence with a superintelligence is compatible with human freedom. Insofar as it would remain true of such a machine that, if it wanted to "eat" us, it could, it seems that we would still be subject to its whims, dominated, and thus unfree. Domination exists, according to the republican tradition, where our rulers have the *capacity* to interfere arbitrarily, regardless of whether they are motivated to do so.

What would be required to preserve human freedom is that a Friendly AI *could not* act against humanity's interests—presumably because its design renders it impossible for it to ever want to. It is unclear whether the existence of such hardwired limits on what an AI is capable of desiring is compatible with claiming it to be an agent and, therefore, "genuinely" intelligent. Regardless, it is doubtful that we could impose such limits on a superintelligence even if they are possible (for some discussion, see Bostrom 2014, 185–208). Moreover, locking in the motivations of AI would also increase the risk that things will go wrong as result of the machine's fixed sense of what our interests are deviating from our own (the "paperclip" problem again). What a republican conception of liberty demonstrates, then, is that there is a profound tension between AI's freedom and our own.

4 The "eyeball test"

Pettit's work provides us with one more conceptual tool to investigate the relationship between power, freedom, and status, which further highlights the tension between the power of super-intelligent AI and human freedom. In a republic, citizens meet as equals of a certain sort. Even if some are wealthy and some are poor, no citizen dominates another. Knowing that they are secure from the arbitrary exercise of power by others, citizens do not need to bow and scrape to their "superiors". They can look each other in the eye. The freedom of citizens is reflected in their *status* as citizens—and vice versa (Pettit 1997, 71–73).

Pettit calls the question as to whether people look each other in the eye in their daily encounters with each other the "eyeball test" (Pettit 2012, 84). The eyeball test tells us whether people feel dominated and—on the assumption that people tend to have an accurate sense of their relationships with others—whether they *are* dominated.

AI is unlikely to have eyeballs. Looking into its—no doubt ubiquitous—CCTV cameras is unlikely to reassure us either that we are its equals or that it thinks of us as such. Indeed, we cannot have a relationship of equals with a superintelligence, because we will not be its equals. The power that even a friendly superintelligence would have over us means that we would effectively still be its pets. Consequently, we will have a strong incentive to pander to it. Just as dogs and cats work to please us so we will feed them, in the future so too may humanity work to please AI to reduce the chance that it might turn against us. Knowing that our luck might run out, we may live in fear that it will. Even if we do not, we will know that we survive only at these machines' sufferance. To exist in such a state is both psychologically demanding and detrimental to freedom by virtue of the increased difficulty of making plans while worrying how one's superiors will respond to them. More fundamentally, though, it is a sure sign of domination and, thus, the absence of freedom.

5 Discussion

Our brief excursion into republican political philosophy, then, has a clear lesson. Solving the Friendly AI problem would not change the fact that the advent of super-intelligent AI would be disastrous for humanity by virtue of its implications for our freedom. Benevolence is not enough to justify making ourselves hostages to the decisions of AI. The pets of kind owners are still pets. As long as AI has the power to interfere in humanity's choices and the capacity to do so without reference to our interests, then it will dominate us and thereby render us unfree. The future Minsky imagines is a dystopian one, regardless of whether AI is friendly or not, and a concern for human freedom provides a strong reason to avoid it.

That "AI's pet" is an unappealing status may seem obvious to some readers. Anyone who has been in a situation where another person has the power to mess with you at their whim—which includes, in particular, many women and many employees—will be able to recognise the ways in which such relationships require that one strive to anticipate the desires, and avoid provoking the displeasure, of the powerful party.

It is, therefore, worth pausing for a moment to speculate as to why so much attention has been—and is being—paid to the question as to how we could make sure that AI is friendly and so little to the fact that the power relationship that would exist between us and super-intelligent AI would itself be enough to render us its slaves. In part, one suspects this is because of the hold that a doctrine of "negative liberty" has over our culture. To be free according to advocates of negative liberty requires only that no-one prevents us from doing what we want—not that no-one could do so if they wanted to.[2] This doctrine is blind to—indeed arguably wilfully obscures—the impact that inequalities of power have on liberty. However, one also suspects that a commitment to a doctrine of negative liberty is also more than a little self-serving on the part of many of those writing about Friendly AI, who are funded by organisations that already have a vast amount of power over those who use their products.[3] To admit that freedom requires political equality and that even a benevolent dictatorship is still a dictatorship would be to call into question the extent to which the wealth and political influence of the institutions funding Friendly AI research is itself compatible with liberty.

The culture, and the material interests, of the corporations working to create AI arguably also play a role in shaping discussion of the "benefits" of AI and thus—although the trade-off is seldom explicitly recognised—whether they might justify our giving up our freedom. The literature on super-intelligence routinely lists fabulous benefits that a friendly AI would supposedly secure for us, which, it is suggested, may justify the—existential—risks involved in pursuing AI (Russell 2019, 98–102). That the price of securing these benefits would be our freedom suggests that they would need to be spectacular indeed. Again, the idea that securing material benefits from one's intellectual superiors might require trading off liberty coheres all-too-nicely with the essentially technocratic framing of the Friendly AI problem. This is not to deny that there might be some circumstances in which we should agree to become AI's pet, but this prospect must be recognised for what it is: threat rather than promise.

Some of the cleverest people in the world are working to realise AI. Some of them declare that there is a non-trivial risk that this project will lead to the creation of entities that might—in the *best* case—relate to us in the way we relate to cats and dogs. Given the difficulties of evaluating claims about progress in AI for the non-specialist, and the quality of the intellects on both sides of the debate about the likelihood this will occur, I struggle to form an opinion on the matter myself. What I *am* confident of is that—even if it were available—a guarantee that our future robot overlords will be benevolent is cold comfort. Worrying about how to create "friendly" AI is a distraction. If we really think that there is a risk that research on AI will lead to the emergence of a superintelligence then we need to think again about the wisdom of researching AI at all.

Acknowledgements I was originally inspired to address this topic after a conversation with Andrew Roberts of the University of Melbourne. Thanks are also owed to Graham Oppy for discussions that have improved the paper. The author is an Associate Investigator in the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (CE200100005) and also contributed to this paper in that role.

[2] The most famous exposition of the idea of negative liberty occurs in the course of Isaiah Berlin's *Two concepts of liberty*. Oxford: Clarendon Press, 1958.

[3] The Machine Intelligence Research Institute (formerly the Singularity Institute), which has driven much on the discussion of the Friendly AI problem is funded, in part, by the Thiel Foundation, a private foundation funded by PayPal co-founder and venture capitalist Peter Thiel. Kurzweil was employed by Google late in his career. The Future of Life Institute, which is highly active in debates about the existential risk posed by AI, was co-founded by Jaan Tallinn, one of the inventors of Skype, and receives funding from Elon Musk, founder of Tesla.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest None.

References

Adams T (2021) Daniel Kahneman: 'Clearly AI is going to win. How people are going to adjust is a fascinating problem'. The Guardian (Online), May 16, 2021. https://www.theguardian.com/books/2021/may/16/daniel-kahneman-clearly-ai-is-going-to-win-how-people-are-going-to-adjust-is-a-fascinating-problem-thinking-fast-and-slow. Accessed 3 May 2022

Anthony A (2017) Max Tegmark: 'Machines taking control doesn't have to be a bad thing'. The Guardian (Online), September 16, 2017. https://www.theguardian.com/technology/2017/sep/16/ai-will-superintelligent-computers-replace-us-robots-max-tegmark-life-3-0. Accessed 3 May 2022

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Chalmers DJ (2010) The singularity: a philosophical analysis. J Consciousness Stud 17(9–10):7–65

Darrach B (1970) Meet Shaky, the first electronic person: the fascinating and fearsome reality of a machine with a mind of its own. Life Mag 69(21):58B–68B (Time Inc., New York)

Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30:411–437. https://doi.org/10.1007/s11023-020-09539-2

Good IJ (1966) Speculations concerning the first ultraintelligent machine. In: Alt FL, Rubinoff M (eds) Advances in computers, vol 6. Academic Press, New York, London, pp 31–88

Kurzweil R (2000) The age of spiritual machines: when computers exceed human intelligence. Penguin Books, New York

Kurzweil R (2005) The Singularity is near: when humans transcend biology. Viking, New York

Larson EJ (2021) The myth of artificial intelligence: why computers can't think the way we do. The Belknap Press of Harvard University Press, Cambridge

Moravec H (1988) Mind children: the future of robot and human intelligence. Harvard University Press, Cambridge

Ord T (2020) The precipice: existential risk and the future of humanity. Bloomsbury, London

Pettit P (1997) Republicanism: a theory of freedom and government. Clarendon Press, Oxford

Pettit P (2001) A theory of freedom: from the psychology to the politics of agency. Polity Press, Cambridge

Pettit P (2012) On the people's terms: a republican theory and model of democracy. Cambridge University Press, Cambridge, New York

Russell S (2019) Human compatible: AI and the problem of control. Allen Lane, Bristol

Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom N, Cirkovic MM (eds) Global catastrophic risks. Oxford University Press, Oxford, pp 308–345

Yudkowsky E (2001) Creating Friendly AI 1.0: the analysis and design of benevolent goal architectures. The Singularity Institute, San Francisco, CA. http://intelligence.org/files/CFAI.pdf. Accessed 11 July 2022