AI & SOCIETY (2020) 35:991–993
https://doi.org/10.1007/s00146-020-00953-9

OPEN FORUM

The rise of artificial intelligence and the crisis of moral passivity

Berman Chan[1]

Received: 10 January 2020 / Accepted: 14 February 2020 / Published online: 4 March 2020

Abstract
In "The rise of the robots and the crisis of moral patiency", John Danaher argues that the rise of AI and robots will dramatically suppress our moral agency and encourage the expression of moral passivity. This discussion note argues that Danaher needs to strengthen his argument by supporting two key assumptions, that (a) AI will otherwise be friendly or neutral (instead of simply destroying humans), and that (b) humans will largely succumb to the temptation of over-relying upon AI for motivation and decision-making in their personal lives.

Keywords Robotics · Artificial intelligence · Technological unemployment · Moral agents · Moral patients · Friendly AI

It is an understatement to say that artificial intelligence is being used more and more in our world. Even though artificial intelligence technology has not reached anywhere near its full potential, several writers look to the distant future and envision all sorts of different impacts on humans. John Danaher (2019a) discusses one particular impact upon humanity, defending the thesis that the rise of AI and robots will "suppress our moral agency and increase the expression of moral patiency" (2019a, p. 133). Danaher argues for this by contending that the dramatic increase in our moral patiency will come as the result of AI's intrusion into three significant human arenas for moral agency: (1) the workplace and employment, (2) political, legal, and bureaucratic decision-making, and (3) leisure and personal activities. Intrusion into these arenas corresponds to three trends coinciding with the rise of the robots. I will argue, however, that Danaher's argument based upon the first trend needs to provide more support for the assumption that robots will pursue a certain kind of takeover (i.e. friendly or neutral) so as not to provoke human resistance and the exercise of our moral agency. Moreover, I argue that Danaher's argument based on the third trend needs to provide an argument for the important assumption that most people will succumb to the temptation of overly relying upon AI for personal motivation and decision-making.

Let us briefly discuss some terminology. Danaher distinguishes robots, which are systems capable of acting in the world, from AI, which need not have this capability. Instead, the latter are typically confined to a computer which assists or offers instructions to humans (2019a, p. 130). However, Danaher uses the two terms interchangeably (so have I above, granting him the connection) and often refers to both at the same time as he argues that the crisis of moral patiency will result from our increased reliance upon both. A moral patient is "a being who possesses some moral status […] but who does not take ownership over the moral content of its own existence" (2019a, p. 132). In contrast, a moral agent is any being who is capable of taking ownership and moral responsibility for its own actions. Danaher is clear that he does not claim that the rise of the robots will result in humans losing their status of moral agency and taking on that of moral patients. Instead, his Thesis is that it will "suppress" the expression of agency-like qualities and "increase" that of patiency-like ones (2019a, p. 133), to such a dramatic extent and affecting the majority of the population that he calls it a "crisis of moral patiency" having "broad civilization-level significance" (2019a, 129). Those exhibiting agency-like qualities take care to gain moral insight in addition to planning and executing action to improve the world or their lives. Those exhibiting patiency-like ones have less moral understanding and ability to act in the world, though as moral patients they can feel pain and have their interests thwarted (2019a, p. 132). Since patiency-like qualities are essentially passive qualities with respect to moral life, I will also refer to moral patiency as moral passivity.

✉ Berman Chan
chan127@purdue.edu

[1] Department of Philosophy, Purdue University, 100 N University St., West Lafayette, IN 47907, USA

With this basic terminology clarified, let us turn to Danaher's argument based on the first of three trends identified earlier. Here, Danaher cites many other writers who argue that robots will replace nearly all human workers, ushering in an age of massive human unemployment (Danaher 2019a, p. 134). Not only this, but his insight here is that this event would deprive humans of a crucial arena for expressing moral agency, as work (employment) provides for society's needs, is how we provide for our families, serves as a place to develop moral virtue, and can become a source of personal meaning.

Now, if the rise of the robots would render us all unemployed, and if this intrusion by AI were part of an otherwise friendly takeover [including even the giving of "the basic income" to each human (2019a, p. 134)] one should agree with Danaher that this would be detrimental to our expression of moral agency. However, there is debate in the literature about whether the takeover would be friendly. Danaher himself (2019a, p. 129) lists some writers who believe it would be friendly, but also one who thinks AI could become an existential threat to us (Bostrom 2014). Elsewhere, Bostrom and a co-author argue that despite the risk of such an unfriendly takeover, we can take steps to ensure AI will be friendly towards us (Muehlhauser and Bostrom 2014), while others are very skeptical and argue that all we can do is hope for the best (Boyles and Joaquin 2019). Danaher wishes to bracket this friendly/unfriendly AI question by envisioning a robotic takeover involving "less fanciful speculation about the likely future intelligence or power of intelligent machines" (Danaher 2019a, p.129). However, I argue he cannot so easily bracket this debate because even though they have lower levels of competency, Danaher is envisioning robots that can completely replace humans in the workforce. What this implies, is that these AIs will be capable of replacing human police officers and soldiers, and so the robots would certainly be capable of attacking humans. Now, if these AIs are indeed friendly or even neutral towards us, then perhaps human unemployment due to robotic takeover will result in moral passivity in the way Danaher argues. Conversely, if the AIs are unfriendly to humans, there is a great likelihood they would seek to destroy humans, rather than merely take away our jobs while providing us with income. This is because humans would be useless to AI and robots, which would be equally effective but more efficient (as Danaher himself grants in [2019a, p. 134]), and also humans would just compete for valuable resources. If this is true, then the ensuing hostilities against humans would provoke conflict which would spur us to action and to exercising moral agency. (If the robots succeed in destroying humans, then the problem would be annihilation instead of moral patiency.) Now perhaps Danaher would disagree that competition for resources would result in AI becoming unfriendly, or even dispute there would be significant competition.

Or perhaps Danaher would argue that the AIs could not become willing to be unfriendly towards humans. However, this could be in tension with these AIs being intelligent and creative enough to replace all human workers, including managers and leaders who make political, legal, and bureaucratic decisions (2019a, p. 135). For this ability might include the capacity to consider all sorts of different alternative actions and also alternative aims. So, there might arise the not unfamiliar tradeoff between the intelligence of these AIs, and the potential that they will be unfriendly towards humans. In other words, if they lack the potential to become unfriendly, then they may not be intelligent enough to bring about complete human unemployment. If instead they are intelligent enough, then the risk of them becoming unfriendly increases. Perhaps Danaher has an argument to show that this is wrong and that the "tradeoff" does not exist in this case; maybe by offering some reason why the alternative aims these AIs can consider would be such that being unfriendly to us is off-limits (without compromising the intelligence and creativity needed to replace all human employment). Whatever the case, Danaher needs to engage more with this debate about friendly vs unfriendly AI and strengthen his case by offering an argument that these AIs will be neutral or friendly in the way he envisions.

Even granting Danaher the argument regarding the first trend of human unemployment, I argue that his argument regarding the third and last trend is in need of further support. To give context to his reasoning there, let us examine his overall argument strategy: first, he has argued that due to the first trend, we will be deprived of moral agency otherwise found in employment. Second, after being shut out of the labour force, we might think that we can nonetheless exercise our moral agency in some other remaining arena–legal, political, or bureaucratic decision-making. But due to the second trend, that machine-learning algorithms will do this decision-making for us, this arena is no longer available to humans (2019a, p. 135). Third, the final and only remaining arena we might turn to is to exercise moral agency in our personal lives and relationships: "We could pursue the humanistic and intellectual pleasures; enhance our personal fitness and well-being; seek out meaningful relationships with others; and produce works of artistic beauty." (2019a, p. 135) However, Danaher argues that even here our moral agency will be all but eroded because of a third trend. This trend is that the rise of the robots would provide a host of AI personal assistants that would not only make our personal decisions for us but also supply the motivation to follow-through with action by cajoling and rewarding us (2019a, p. 135). So, in a future world of highly developed

AI algorithms that can do all this, humans would rarely: make their own decisions, deepen in understanding of reasons for moral action, or cultivate their own character traits of courage and perseverance. Danaher envisions that humans will lead carefree lives, and he even muses that the likely emergence of "pleasure bots" would arguably give humans very pleasurable lives. But he is understandably alarmed that humans would not be exercising moral agency.

Danaher's argument from this third trend needs to be strengthened, however, because it assumes that the majority of people will largely succumb to the temptation of over-relying on these AI personal assistants. Earlier in his article, when he discusses sex robots and pleasure bots in general, Danaher appears to either assume or state as a possibility that most people would succumb to AI, constituting a "civilization-level threat", because of the pleasure and benefits they provide: "[H]umans might become increasingly passive recipients of the benefits that technology bestows […]. [T]he subtle way in which they play upon our psychological biases and temptations may be the problem" (2019a, p. 131). However, as this claim is either an assumption or expressed merely as a possibility, Danaher should provide some empirical or other reason to hold it. Otherwise, he would only have shown that the temptation exists, but not that most people would succumb to it to create the crisis of moral patiency. Thus, Danaher has not said anything to rule out that most people would likely exercise their moral agency by deciding not to over-rely on AI assistants in the first place. This sort of counterargument gains traction because the trend under discussion implies that AI assistants would be maximally pervasive and intrusive in our personal lives, providing almost all of our decision-making and motivation, and thus it is reasonable to question that most people would allow it.

This finds further support in Robert Nozick's classic thought experiment about the experience machine: "Suppose there were an experience machine that would give you any experience you desired […] Should you plug into this machine for life, preprogramming your life's experience?" (Nozick 1974, pp. 42–43) This thought experiment may constitute an argument against the view that pleasure is all that matters in life, but that is not my focus here. What is salient is that the example strikes a chord with many of its readers, leading them to conclude due to several reasons that they would not plug in (here I am focusing on the question of would they, not even should they, plug in). Many want to be living their own lives autonomously instead of one run by the experience machine. Also, they want to be a certain person, with certain real character traits. These are what many readers want even though the experience machine would provide them amazing pleasure and would also apply to why many people would reject over-reliance on pleasure bots or (convenient) AI personal assistants. Now, it is true that not all of Nozick's readers would think this way, but at least some significant proportion would, and so this counts in favour of people not succumbing to the temptation relevant to the third trend Danaher identifies. Conversely, it is also true that despite most people's intentions of avoiding becoming maximally dominated by AI personal assistants, they might nevertheless succumb to temptation when they are tested.[1] However, the burden falls upon Danaher to demonstrate from empirical or other arguments that in spite of these intentions most of them will succumb, as he is advancing the positive argument that most people will become morally passive with the rise of the robots. For similar reasons, with regards to the first trend of unemployment, the onus is on Danaher to provide the needed support to show that the rise of the robots will be otherwise friendly (or neutral) to us in the first place.

References

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

Boyles RJM, Joaquin JJ (2019) Why friendly AIs won't be that friendly: a friendly reply to Muehlhauser and Bostrom. AI Soc. https://doi.org/10.1007/s00146-019-00903-0

Danaher J (2019a) The rise of the robots and the crisis of moral patiency. AI Soc 34(1):129–136

Danaher J (2019b) Automation and utopia: human flourishing in a world without work. HUP, Cambridge

Muehlhauser L, Bostrom N (2014) Why we need friendly AI. Think 13(36):41–47

Nozick R (1974) Anarchy, state, and utopia (Vol. 5038). Basic Books, New York

[1] In Danaher's earlier Automation and Utopia (2019b), in which he also writes about the crisis of human passivity, he might have provided some theoretical resources that could be used to argue that humans will by and large succumb, despite my objections. If so, it would have been helpful for Danaher to deploy them in his article and clearly spell out how he is arguing that humans will succumb to the temptation.