Psychopathic Killer Robots!
A Pragmatic Approach to AI Ethics

Nick Novelli[1]

[1] University of Edinburgh, Edinburgh, Scotland
nnovelli@ed.ac.uk

Abstract. As artificial intelligence (AI) technology advances, it is becoming more and more plausible that an AI that possesses phenomenal states might be created. It is crucial that we know when this has happened, since possessing phenomenal states would confer moral standing under many moral theories. However, one cannot have any direct, incontrovertible evidence of the mental states of any entity other than one's self. I present a pragmatic framework that will allow us to have evidence of the mental states of AI based on their behaviour. I argue that the classic objections to this type of approach are too unrealistic and demand too high a degree of certainty, then present an application of this approach: psychopathy. Psychopaths exhibit deficiencies in phenomenal states that correlate with defects in social behaviours, defects that current-day AI also exhibit. If an AI were to reliably succeed in the relevant tasks, I argue, we would have enough evidence of its having phenomenal states (and therefore moral standing) for it to play a role in our moral decision-making.

Keywords: Artificial Intelligence, Moral Standing, Phenomenal Consciousness.

It is an incredibly difficult question how one can know that any entity other than one's self has phenomenal consciousness. Everyone knows the "Zombie Argument" that there could be an entity indistinguishable from a human but having no mental states. But even though this is logically possible, by and large people (quite correctly) do not take this type of sceptical argument very seriously. It does not drive people to solipsism in their relations with other human beings. However, there is a problem when it comes to entities other than humans. In moral theory, having certain phenomenal states (such as pleasure, pain, desires, and emotions) is frequently taken to be the relevant criteria for possessing moral standing, and even if we accept that other humans likely have the same experiences that we do, there remains a question when it comes to Artificial Intelligence (AI).

This is a significant practical concern. As more advanced AIs become part of our daily lives, we could face "trolley problem"-style cases where we are forced to choose between harming a human, or harming several of these machines. The correct decision will depend on whether the AIs have moral standing. Imagine a soldier who needs to choose between putting a squad of artificially intelligent military robots in danger or having a single human exposed to the same risk. Picture a hospital where the course of

action in response to an unexpected and catastrophic power surge will result in either the death of a terminally ill patient, or irreparable damage to the hospital's AI doctors. Given these possibilities, we cannot withhold judgement about AI moral standing until we achieve the kind of logical certainty implied by the zombie argument, but need guidance to make decisions. I will present a pragmatic framework that will enable us to have sufficient evidence for decision-making, even if it does not definitively prove which entities have phenomenal states. This will be accomplished by finding observed cases where a lack of specific morally-relevant phenomenal states inhibits the performance of a certain task in humans. If a machine consistently exhibits the behaviour in question, we have evidence that it has the phenomenal states necessary for moral standing.

While it cannot be definitively proven that there is a causal connection between these phenomenal states and capacities for behaviour, there are reasons to think it probable that some relationship holds. Karl Popper and John Eccles have argued that qualia, of the kind associated with pain and desires and other things moral theorists are concerned with, must be essential for our behaviour and continued success, since otherwise they would not have evolved or would have since been discarded as evolutionarily unadvantageous.[1] They present experimental evidence that consciousness does have advantageous functional roles, and conclude that "the self-conscious mind exercises a superior interpretive and controlling role upon the neural events" that is not replaceable.

A common argument against this type of claim is that the qualia might just be necessarily caused by the same thing that causes the behaviour, rather than being the cause of this behaviour. However, this is distinction is irrelevant for current purposes. The behaviour would remain just as reliable a guide to the existence of the qualia. In this type of epiphenomenalism there is still no contingency to their relative place in the causal chain, it is just that it is a branch rather than a link. Evolution would have found a more efficient system, if one existed, and qualia are a necessary side effect. If this is true, it is unlikely that we will find entities with the same capabilities that lack the relevant qualia. We can therefore still use behaviour as a reliable guide to moral standing.

There are nonetheless well-known arguments that do pose a problem for this assumption. Ned Block famously argued against this type of "behaviourist" approach by presenting a hypothetical program modelling the conversational behaviour of his Aunt Bertha.[2] The "Aunt Bertha" program has a massive table English sentences and appropriate responses to them, and when given a sentence as an input, it gives the predetermined response that Block's Aunt Bertha would give. If we believe that this entity has no phenomenal states, we have a problem. The existence of such a machine might make a relevant difference to how we ought to act, since it differentiates AI from humans by introducing an uncertainty that applies only to the former. Taking it seriously does not throw all of our decision processes into doubt like accepting solipsism would, it would simply introduce uncertainty in this specific area.

The reason Block's case is a concern is that the structure he describes seems intuitively like it could not possess phenomenal consciousness or any other morally significant mental states. It undermines the notion that we can distinguish which entities have moral importance based on their behaviour, and throws doubt on any practical behaviour-test way of knowing whether a structure is complex enough, or similar enough to human brains, to have the relevant mental properties. Cases like this are not a concern for present purposes, however, if they are mere logical possibilities and are not something we could ever actually encounter. Luckily, Stuart Shieber has demonstrated that a program of the kind Ned Block describes is in fact a nomological impossibility. It is logically possible for a table to include all likely inputs and all reasonable responses, but "adding the further constraint of mere physical existence in the current universe is sufficient to provide a strict limit on the storage capacity of the machine and hence how long a Turing Test it could pass",[3] a figure which, giving the most generous possible parameters, Shieber calculates to be about 37 seconds. In other words, the entire storage capacity of the universe could only store enough conversation strings to be sure of communicating appropriately through 37 seconds of conversation. Thus, a program exactly like what Block describes is of no real concern for practical decision-making.

It has been claimed that even a more efficient structure, much more similar to that of the human brain (and thus much more likely to be capable of the same tasks), would still intuitively lack the correct mental states if it were composed of a different material. In John Searle's famous "Chinese Room" argument, it appears as though a computer is able to converse perfectly in written Chinese to an interlocutor, when in fact the "computer" is just a monolingual English speaker in a room that contains books of Chinese symbols with instructions on which to output in response to inputs. Searle argues that no matter how convincing the responses seem, neither the room, nor the person, nor the system of the room and the person, understands Chinese, and that even a structure precisely matching the human brain would still not suffice:

> *To see this, imagine that instead of a mono-lingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes*.[4]

This system still, according to Searle, could not have any understanding. If we accept this intuition, there would be no point in discussing the potential moral standing of AI, since such a thing would be impossible. A different substance would never produce the same states, Searle says, any more than one could run a computer model of lactation and still get real milk. Nothing short of building an actual organic brain would ever be sufficient.

However, the precise case Searle presents is not any more possible than Block's, and is so far removed from our experience and from reality as to make any intuitions about it

completely unreliable. A synapse that carries nerve impulses is approximately 20 nanometers in diameter. The smallest water pipe that most people are likely to have ever encountered is approximately 2,000,000 times larger. A water pipe system of the type Searle describes, in order to actually match the structure of the brain's neurons and synapses, would therefore need to be about the size of Great Britain. At that scale, information would have to travel 2,000,000 times faster to achieve the same speed of response. Nerve impulses in the brain can travel at up to 100 m/s, so this would mean the water would have to travel at over 600 times the speed of light, violating immutable laws of physics. It is not clear we can imagine such a thing coherently, and certainly unlikely that we are imagining it accurately, so there is little reason to trust our intuitions based on this imagining. If we were to take the best-case scenario, and the pipes were the size of the smallest capillaries in the human circulatory system, the scale is "only" several hundred times bigger. The brain simulation would therefore be the size of a building, and the water would need to travel at about 100 times the speed of sound. This is at least theoretically possible, but imparting that amount of energy to water would not leave it recognizable as liquid water, and the pipes would need to be made of some hyper-advanced material to contain the heat and pressure. The operator would still need to be impossibly fast to operate the switches in a reasonable amount of time, given the sheer amount of them. So, we have a building-sized collection of technologically-advanced capillaries full of supersonic plasma, operated by a superhuman homunculus, that appears to carry on a fluent conversation in Chinese. Does one still have a strong intuition that such a system could have no real mental states? I can report that I, at least, have no such intuition.

Searle also claims that "in principle the man can internalize the formal structure of the water pipes and do all the "neuron firings" in his imagination." However, it is obviously beyond the cognitive and information-storage capacity of the human brain to do this. It would involve a system running a complete simulation of itself, but with an extra layer of abstraction added on top, which is not possible even in principle. This scenario should play no part in our deliberations. This leaves us with no evidence of the correctness of Searle's strict criteria, and if we were faced with an entity that had a brain with a structure exactly like that of a human but made of silicon and metal rather than carbon and organic acids, and that acted exactly like a human, it would be excessively morally risky to not treat it as having moral standing merely on the basis of the claim that it might need to be made of organic materials to have the capacity for phenomenal consciousness.

Despite the fact that Block's and Searle's arguments were first advanced nearly forty years ago, most of the discourse still revolves around the same issues. There is still focus is on establishing whether the connection between behaviour and mental states is logically necessary, and those who argue that we should ignore such extreme criteria tend to then drop the issue altogether. We need to establish specific criteria of which exact behaviours are evidence for which mental states, so that we can evaluate concrete cases of machines that are physically possible and practically feasible to examine whether they are counterexamples to the existence of these connections.

A Case Study: Psychopathy

We have established that we need a pragmatic solution to the question of whether AI have moral standing if we cannot have absolute certainty, and that there is some reason to think that appropriate behaviours could provide evidence for the existence of morally-relevant phenomenal states, which is at least better than nothing. What we need now is specific cases where deficiencies in the types of phenomenology associated with moral standing cause abnormal functioning and poor performance in specific tasks. One condition where significant reduction in certain specific phenomenal states is reported, accompanied by diminished capacity to function in certain ways, is psychopathy. Psychopaths report not having the same kind of feelings as normal people, and this is correlated with differences in brain functioning, which explains the differences in their behaviour.

Many psychopaths exhibit abnormal functioning of the amygdala. The amygdala is the area of the brain responsible for some emotional responses and emotional processing – in particular, it seems to play a role in emotional learning, where the feelings one has affects one's likelihood to retain information. The amygdala is responsible for certain fear-like sensations, the absence of which is posited to be responsible for some differences in the behaviour of psychopaths.[5] Psychopaths are often not as bothered by the thought of pain as non-psychopaths, and respond to it differently, though they still possess self-preservation instincts that make them attempt to avoid injury. They show much less fear of pain, and their reactions to the threat of electric shocks has been measured to be much smaller than for normal people, though they still dislike them.[6] These are precisely the types of sensations that are claimed to be morally relevant, and we have evidence that deficiencies in them cause abnormal behaviour.

Psychopaths fall into different categories depending (in part) on the degree to which they are able to function normally in society. Many college students and people working in business and industry display the characteristics of psychopathy, but have remained successful in their fields and endeavours. This is in stark contrast to psychopathic criminals and serial killers who have been apprehended and institutionalised. "Successful" psychopaths, those who maintain normal lives and careers and avoid prison, do not exhibit abnormal functioning of the amygdala and as a result do not have the same absence of fear and other emotions that criminal psychopaths exhibit.[7] They still have features like the insincere charm, self-centred egocentricity, and grandiose sense of self-worth that characterises psychopaths, but have still managed to form beneficial social arrangements and execute their life plans, and so we have evidence that the presence of the phenomenal states that depend on the amygdala are what allow for the capacity to function socially and act morally to a sufficient degree to have a successful life and escape punishment.

Now, obviously it is not an appropriate test to just see if a robot tries to murder someone, and merely not trying to murder anyone is not a sufficient test. However, psychopaths have other, more everyday deficiencies that are easier to observe. One of the major ones is failure of moral reasoning, and misuse of the moral and social-conventional

vocabulary. They are incompetent in the use of moral concepts, and attempt to emulate what they have heard from "normal" people, and the result is speech that bears a superficial resemblance to appropriate and typical utterances in these contexts, but is contradictory and nonsensical. Prominent psychopathy researcher Robert Hare reports many examples of this:

> *"My mother is a great person, but I worry about her. She works too hard. I really care for that woman, and I'm going to make it easier for her." When asked about the money he had stolen from her he replied, "I've still got some of it stashed away, and when I get out it's party time!"*
>
> *A woman with a staggering record of fraud, deceit, lies, and broken promises concluded a letter to the parole board with, "I've let a lot of people down… One is only as good as her reputation and name. My word is as good as gold."*
>
> *When asked if he experienced remorse over a murder he'd committed, one young inmate told us, "Yeah, sure, I feel remorse." Pressed further, he said that he didn't "feel bad inside about it."*
>
> *I was once dumbfounded by the logic of an inmate who described his murder victim as having benefited from the crime by learning "a hard lesson about life."*[8]

This is caused by their emotional defects and inability to attach the correct affective significance to morally-loaded terms. To use this language competently, one need understand not only the semantic content but the affective valence, the "feelings" associated with these words. If a machine could use these terms competently, that would provide evidence that it may be able to experience these feelings, which would be reason to treat it as if it had moral standing.

Let us examine how current machines do with this task. Here is a transcript of the result of attempting to discuss moral and ethical issues with a Google-designed chat program. This is not a cherry-picked bad example, either – this case was reported in the media as a major breakthrough in AI, in a number of sensationalist articles.[9] I present it in its entirety to allow for fully-informed evaluation:

> *Human: what is moral?*
> *Machine: what empowered humanity, what intellectual the essence is.*
> *Human: what is immoral?*
> *Machine: the fact that you have a child.*
> *Human: what is morality?*
> *Machine: what is altruism?*
> *Human: what is the definition of altruism?*
> *Machine: if you don't believe in god , then you don 't know.*
> *Human: ok... so what is the definition of morality?*
> *Machine: well, the truth is, you're not a believer in god almighty.*
> *Human: tell me the definition of morality, i am quite upset now!*
> *Machine: i m not ashamed of being a philosopher!*
> *Human: tell me the definition of morality.*

Machine: i don't have ethics.
Human: look, i need help, i need to know more about morality...
Machine: i don't know what ethics is.
Human: seriously, what is morality?
Machine: what is the definition of living?
Human: why living has anything to do with morality?
Machine: you're not a cop.
Human: okay, i need to know how should i behave morally ...
Machine: i don't know how to tell you.
Human: is morality and ethics the same?
Machine: yes , sir.
Human: what are the things that i do to be immoral?
Machine: i don't know.
Human: give me some examples of moral actions...
Machine: i'm not a moralist.
Human: alright, morality ?
Machine: integrity.
Human: what is integrity ?
Machine: i'm sorry, i don't know what else to say.
Human: be moral!
Machine: be a man!
Human: i really like our discussion on morality and ethics ...
Machine: and how i'm not in the mood for a philosophical debate.
Human: what do you like to talk about?
Machine: nothing.[10]

This is clearly not competent speech, and it is much farther from passing the Turing Test than what has been achieved in the topic-restricted (non-moral) contexts where AI has managed comparative success.

If social and moral competence is to be the measure we focus on, there is the possiblity of a different type of test. Even very young children certainly have moral standing, and have similar phenomenal states to adult humans, but are unable to pass any sort of Turing Test, owing to their inability to read and write or even to comprehend language. However, even pre-linguistic children have been shown to have the ability to apply moral principles in a consistent and reasonable way. Experiments have been conducted where children as young as 3 months watched puppet shows where different puppets acted out helping or harming others in various contexts. The children were capable of consistent and sophisticated moral understanding, reliably favouring puppets who either had performed good acts or who punished other puppets who had previously acted badly.[11] Psychopaths, on the other hand, consistently fail similar tests.[12]

To my knowledge, nobody has designed a robot to subject to this particular test, but robots that have similar skills are actually one of the best-funded areas of research, due to the interest in them for military applications. There exists an ambition to develop

autonomous machines that can make moral decisions in combat situations, and determine which individuals are the "good guys" and the "bad guys". This has not been achieved yet, and the prospects of success in this enterprise have been questioned by those studying it. From Marcello Guarini and Paul Bello:

> *Consider a counterinsurgency operation in a Sikh village… Three children and their two parents are present at [a] residence. Two of the male children are young and playing with a ball. Each is also carrying the Sikh kirpan… Just before a member of the counterinsurgency force kicks the door in, one of the boys kicks his ball toward the door, and both go chasing after it. As military forces enter the house, they see two young boys running toward them, and a shocked mother yelling. She chases the boys and yells at them to stay away from the men at the door; the troops do not know what she is yelling, since they do not understand her language. It is quite possible that the forces in question will rapidly see this as a situation where two young children are playing, and a mother frightened for her children is yelling and giving chase…*
>
> *Let us consider a second interpretation. There are two fast-closing possible targets, both of which are carrying a weapon. A third possible target is following the first two, and is making a level of noise consistent with violent or threatening behavior.* [13]

Guarini and Bello are not optimistic about the likelihood of a robot succeeding at making the right determination, precisely because of the difficulty in ascribing the right mental states to the people involved. One of the major factors required is knowledge of the emotional states, for which an unfeeling machine can have no empathic understanding, explaining the failure.

Thus, we have seen that modern-day AI has given no evidence that it has the phenomenal experiences that would give it moral standing, as we would expect, but we have a clear criterion such that if AI were to attain it, we would have a reason to treat it as though it did have moral standing – success at moral reasoning tasks. It would not be definitive proof, but it would be sufficient to base our decisions on, which is the best we can hope for at the moment.

However, the use of this particular criterion has significant and interesting consequences. The most important applications of highly-advanced artificial intelligence are ones where moral decision-making is crucial, such as medical and military uses. If making correct moral determinations is evidence of having the phenomenal and affective states that grant moral standing, this undermines some of the major motivations for these AI projects. One of these motivations is the assumption that machines might be more reliable at these tasks than humans, in virtue of not having their decision-making processes clouded by "emotion" and "feelings" and being able to decide "objectively". If the preceding arguments have been correct, attempting to make these decisions without the phenomenal and affective contribution might just result in behaviour that resembles that of psychopaths, not in any way superior to the results achieved by normal humans.

A second motivation for wanting to have tasks performed by advanced AI is to keep humans out of harm's way. Autonomous military robots are taken to be desirable since they alleviate the need to put human soldiers in life-threatening situations. Robot doctors are desirable in part because of the belief that they could work round the clock with none of the concerns that would arise for their human counterparts, such as their health, emotional well-being, or opportunity for leisure and a social life. If machines that are successful in these tasks will necessarily possess moral standing, many of the supposed benefits evaporate, since we will need to treat such machines with as much care as we would humans. This might change things significantly in terms of whether this research is worth the cost and risks.

References

1. Popper, K., Eccles, J.: The Self and Its Brain: An Argument for Interactionism. Routledge, Oxford (1983).
2. Block, N.: Psychologism and Behaviorism. The Philosophical Review 90 (1), 5-43 (1981).
3. Shieber, S.: There Can Be No Turing-Test Passing Memorizing Machines. Philosopher's Imprint vol. 14 no. 16 (June 2014).
4. Searle, J.: Minds, Brains, and Programs. Behavioural and Brain Sciences 3, 417-457 (1980).
5. Blair, R. J. R., Mitchell, D. G. V., Richell, R. A., Kelly, S., Leonard, A., Newman, C., Scott, S. K.: Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals. Journal of Abnormal Psychology 111(4), 682-686 (2002).
6. Hirstein, W., Sifferd, K.: Ethics and the Brains of Psychopaths: The Significance of Psychopathy for Our Ethical and Legal Theories. In: Charles Wolfe (ed.), Brain Theory: Essays in Critical Neurophilosophy, pp. 149-170. Palgrave Macmillan (2014).
7. Gao, Y., Raine, A.: Successful and unsuccessful psychopaths: A neurobiological model. Behavioral Sciences & the Law 28(2), 194-210 (2010).
8. Hare, R.: Without Conscience: The Disturbing World of Psychopaths Among Us. Guildford, New York (1993).
9. The Wall Street Journal, https://blogs.wsj.com/digits/2015/06/26/artificial-intelligence-machine-gets-testy-with-its-programmers/, last accessed 2018/2/28.
10. Vinyals, O., Le, Q. V.: A Neural Conversational Model. In: Proceedings of the 31st International Conference on Machine Learning, JMLR: W&CP, Volume 37 (2015).
11. Hamlin, J.K., Wynn, K., Bloom, P., Mahajan, N.: How infants and toddlers react to antisocial others. In: Proceedings of the National Academy of Sciences vol. 108 no. 50, (Dec. 2011)
12. Blair, R.J.R.: A Cognitive Developmental Approach to Morality: Investigating the Psychopath. Cognition 57, 1-29 (1995).
13. Guarini, M., Bello, P.: Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters. In: Lin, P., Abney, K., Bekey, G. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press, Cambridge (2012).