Nikhil Garg

The Ethics of Conscious Artificial Intelligence

Introduction

In our lifetimes, a single machine will be able to outperform humans in any given task. This fact will beget a version of the problem of other minds aimed toward those computers that can pass the Turing Test. And just as gaining awareness of the mental life of animals has brought a revolution in the understanding of animal rights, the exponentially-increasing capabilities of technology will force a decision regarding the rights and ethical obligations of intelligent technology. In this paper, I present some criteria for rights, extend and attempt to resolve the problem of other minds in respect to non-biological life,[1] and present the ethical implications[2] of such intelligence. My argument is that resolving the other minds problem in respect to computers is necessary but impossible. Even more so than in the case of animal rights, inevitable technological advances will produce a rights-crisis that will freeze our ability to act as we struggle to determine the ethical implications of using potentially Conscious technology.

Ethical Theories and Rights

Rights Criteria

Before attempting to answer whether technological life has rights, one must first determine the criteria for rights. In an appeal to human intuition, there are two factors involved in the

[1] At the risk of using loaded language, I use the following terms interchangeably: non-biological life, intelligent technology, intelligent computers, and artificially intelligent beings. Here, 'life' or 'being' refers to an entity that *behaviorally* is indistinguishable from a human, a la the Turing Test. When referring to Consciousness, or conscious*, as Professor Strawson defines it (with its qualia and what-is-it-like-to-be-ness), I specify 'Consciousness,' 'Conscious being,' or 'Conscious Technology.'

[2] When I ask whether one has rights, I do so in the broad sense, not in a specific deontological rights based theory sense. I seek to answer whether non-biological life can be viewed as the equal to biological (and, more specifically, human) life, whether in a utilitarian calculus or in a deontological sense. For example, one can talk of fundamental human rights, but still weigh these rights (and rights violations) against each other.

determination[3]: intellectual capability (language abilities, group formation, learning, task completion) and Consciousness (the ability to feel emotions and suffer, personal awareness, ability to experience). Many use the former factor to justify excluding animals and the latter to exclude non-biological life which could pass the Turing Test. Regardless, these factors suggest two distinct but related questions concerning granting rights to intelligent technology: 1) supposing intelligent technology could gain Consciousness, does it then have rights?, and 2) does intelligent technology, without Consciousness, have rights? Note that these questions each have two components: the descriptive component (in a futurist society, would moral intuitions and legal necessities yield rights for such technology), and a prescriptive component (should such technology have rights).

Conscious Non-biological Life

Integration into Ethical Theories

To answer the first question, concerning the rights afforded to Conscious non-biological life, one must first determine, under each ethical theory, what factors into a moral decision and action. Under consequentialist theories, the moral action is one that maximizes (or minimizes) some metric – happiness, utility, suffering, universal intelligence[4]. Conscious technology, given that it can experience happiness and suffering, trivially must be considered in such calculus. Similarly, under deontological ethics, agents that meet some criteria – self agency, autonomy,

[3] There is another component involved: the ultimately intellectually unjustifiable feeling that being biologically human is the most important factor. I suspect that, to some extent, the two factors I mention are just rationalizations explaining that intense feeling. I will deal with this shadow factor later on.

[4] In his book, *The Information: A History, A Theory, A Flood*, James Gleick argues that information underscores the universe and the human goal should be to further the quest toward spreading intelligence throughout the universe, even if it means the extinction of modern day humans.

consciousness, suffering, and/or consciousness of consciousness, depending on the version – have fundamental, inviolable rights, and, as far as Conscious technology can meet those criteria, it deserves those same rights. Under either ethical theory, Conscious technology should thus have rights.

Anthropomorphic Concern

The only barrier to widespread acceptance of these rights is the species-centric view which maintains that as human beings, our ethical responsibility extends only to other humans. I think that this view is more a moral intuition than a rational position, though it is no less powerful as a result. Slavoj Zizek, in *The Sublime Object of Ideology*, presents the too-common situation where "one knows the falsehood very well, one is well aware of a particular interest hidden behind an ideological universality, but still one does not renounce it" (28-30). Consider the case of animal rights, the one most parallel to the coming debate over artificial intelligence rights. Many love their pets and understand the suffering that animals go through in factory farms and slaughterhouses. However, they still may eat meat and use other animal products, while simultaneously aware that the atrocities committed are unconscionable. We ignore inconvenient truths and continue as if we are unaware of them. The danger of Conscious technology, if indeed it is possible, is that society ignores the conclusions of its ethical theories and relies solely on human-centric moral intuitions.

Response to Anthropomorphic Concern

While it remains to be seen whether the march toward a larger recognition of animal rights will prove successful, I think that the recognition of the rights of Conscious technological life is inevitable. In *The Singularity is Near*, Ray Kurzweil argues that before purely non-biological Conscious life is developed, humans will grow comfortable with augmenting their own

capabilities with technology – artificial limbs and organs, brain implants, additional sensors. We will each become our own Ship of Theseus, and, as the biological portion of our body and brain shrinks, humanity will grow more comfortable with non-biological Consciousness, and the intuition on ethics in respect to non-biological life will evolve with us. Conscious technology thus both deserves and would get the same rights as humans if the technology is realized.

Non-Conscious Intelligent Technology

(Non)-Integration into ethical theories

The second question – should intelligent technology, without Consciousness, have rights, and will it? – is easier to answer in theory. The immediate intuition is that such non-biological life, without the ability to *feel* suffering or happiness, should not factor into any utilitarian calculus that measures those qualities. Furthermore, it would not meet the essential criteria for deontological rights/duties (consciousness of some sort). The technology will be able to indistinguishably (from a Conscious being) discuss and express pain, happiness, and discontent (an essential part of passing the Turing Test), but it will not experience it and so, in theory, should not factor in ethically.

Will Artificial Intelligence be Conscious?

Argument from Analogy in the Problem of Other Minds

Thus, to easily answer whether to consider intelligent technology in one's ethical decisions, one must answer – independent of implementations – whether such technology is Conscious. However, as in the human case, the problem of other minds emerges. It is impossible to observe qualia from another person's perspective, and, in the technology case, one cannot just use the

argument from analogy to dismiss the question as overly intellectual. The argument from analogy uses abduction to justify the assumption of the existence of other minds. I can assume that they think and experience qualia as I do because other humans have similar bodies/brain structure, have developed evolutionarily exactly as I have myself, and behave similarly. These explanations do not work in the case of technological life, and one could conceivably argue that the best explanation is that computers simply execute pre-determined algorithms rather than develop their own. Thus, one must seek other mechanisms to answer whether intelligent computers are Conscious.

An Attempt to Say Yes: Turing Test

One such mechanism is the Turing Test, which uses a behavioral definition of consciousness. For Alan Turing and other behaviorists, the consciousness factor reduces to the intellectual capability factor – computers become conscious when they become indistinguishable from a Conscious being (a human being). Regardless of whether Turing meant the test to be an ontological test or an epistemological test, it cannot be ontological. It replaces the question of whether computers can be Conscious to whether one can know that the computer is not Conscious. The Turing Test thus becomes a weaker version of the argument from analogy – without the first link of similar bodies. It relies solely on behavior and is even more unsatisfactory than the original argument from analogy, which at least is the best explanation. Thus, though a necessary condition to be Conscious, behaving indistinguishably from a human is not sufficient. Given the long, inconclusive debate on the problem of other minds in respect to other humans, it's unlikely that an immediate solution will be found proving that computers are Conscious.

An Attempt to Say No: The Chinese Room

Formulation and History

One could then attempt to prove the inverse, that intelligent technology cannot become Conscious and thus would not deserve rights. John Searle attempts such an argument through his Chinese Room thought experiment, which illustrates the intuition that even a computer that passes the Turing Test *cannot possibly* contain the essence of human Consciousness. This argument has a long history in philosophy, dating back to at least Leibniz, who argued against materialism by imagining increasing the size of brain to that of a mill and then walking through to find the color red[5]. These experiments attempt to circumvent an inability to peek through another's perspective by appealing to the intuition that pieces of silicon cannot think as I do.

A Response from Kurzweil: An attack on intuition

However, as Ray Kurzweil argues, human intuition simply cannot work at the level of complexity needed to build a computer that can pass the Turing Test (and thus be the type of intelligent machine I discuss in the paper). Computers that could pass the Turing Test would have a design and software much more complex than presently imaginable – one possible solution often touted is to simply simulate every neuron and connection in the brain once the computing power is available. The natural, strong intuition presented by Leibniz and again by Searle is thus inadequate given complex enough technology.

Turing Machines and the Recovery of Intuition

Kurzweil's attack on the human intuition can be fully answered, funnily[6] enough, by another one of Turing's formulations: the Turing machine. In computer architecture circles, the idea is that

[5] In this sentence, I assume that computing Consciousness has the necessary condition of materialism being true. I was about to write my paper on such conditions (necessary and/or sufficient), but decided not to because of a lack of interesting things to say.

[6] Or at least I find it funny.

any Turing machine (which all modern computers are), given enough time and memory, can complete any task that any other Turing machine can complete[7]. However, combined with the Turing Test, this fact would require that either today's computers are Conscious or that Consciousness emerges from time and/or memory. From my intuition, the former proposition is impossible. Today's machines simply execute written instructions. The second issue is plausible but unlikely: emergence would suggest that either time (an external condition) gives an object Consciousness or that, by adding enough RAM sticks to my computer, I could make it experience the color red! Though I do not believe in such emergence, I cannot claim to know[8].

Resolution: No one knows (or can know)

Thus, after Turing, Searle, Leibniz, Kurzweil, Turing again, and others, the question – that of the Consciousness of computers – settles into the familiar one of emergence. Can Consciousness emerge from non-Conscious technology? No definitive research exists that could answer the question[9], and the technology to make a serious attempt at general intelligence has not yet been developed. To complicate this question, the computer that passes the Turing Test may not be the binary logic computer of today, but rather a neural machine or a quantum computer. Regardless, there is no easy answer to the question of whether intelligent machines are Conscious.

The Ethical Quandary

[7] Turing demonstrated the concept by writing a program to play chess and, lacking a computer which could run it, simulating the program himself.
[8] Adding memory potentially reopens Kurzweil's criticism, but that's where my willingness to give credence to the criticism falters.
[9] I think Wittgenstein was on to something when he said that one should not speak of something that one cannot speak of intelligently. I consider the entire Panpsychism/emergence debate to be non-verifiable, non-falsifiable, metaphysical nonsense.

This fact raises a large ethical quandary: a moral agent would need to somehow distinguish between Conscious non-biological life and intelligent non-biological life. This is impossible as it would require peeking through the computer's perspective. Unfortunately, the decision is both forced and momentous[10], in the terminology of William James. Either we consider intelligent technology in our ethical decision making, or we do not. Do the former, and humans (and other Conscious life) may suffer as undue weight and reverence is given to non-Conscious technology. For example, society may not be able to take advantage of the numerous benefits of artificial intelligence if it adhered by the rule to not treat it solely as a means, or if one must consider the potential 'happiness' of a piece of silicon before acting. On the other hand, if society chooses the latter approach and uses future technology without limits, then we could potentially be ignoring the suffering of trillions of beings, much as we do with animals in the status quo. Either way, and under any rational ethical framework, humans face moral failure, a risk that cannot be mitigated.

Conclusion

After much interdisciplinary research and attempts to understand our technological future, humans must accept a fundamental, and depressing, truth. The coming century will bring advances that will revolutionize every aspect of our lives, at first for the better. But, inevitably, it will give new importance to philosophical questions centuries old, questions that, up to now, have been little more than ivory tower pursuits[11]. Answering questions such as the problem of other minds in respect to intelligent technology will be necessary to live a fulfilling, ethical life,

[10] I do not necessarily mean so in the religious sense.
[11] I'm not saying that questions, in the philosophy of mind for example, have been useless, but rather, they have not been necessary to have a full, ethical life.

but the questions will prove impossible. Humanity will be forever stuck, either not doing enough for its own or potentially leading a new era of slavery of Conscious beings.