Interstellar Might Depict AI Slavery

Keith Wiley

Put aside your morals momentarily — and now consider the perfect slave. What makes him such an excellent slave? We can circuitously list indirect qualities, say obedience, an unrelenting worth ethic, total self-disregard, absolute willingness to self-sacrifice, etc. But there is a more direct consolidation of these ideas: a lack of self-determination, of even the notion that he could hypothetically possess self-determination. The perfect slave isn't restrained under chain and whip. These are extraneous and wasteful costs to the slave's owner (not just in materials but in the man hours of the guards).

A slave owner doesn't explicitly desire such depravities, but rather relies on them to keep less perfect (but more realistic) slaves in check. To own the perfect slave imposes none of these extra burdens. Let us borrow an anachronistic term and describe the perfect slave as "lobotomized". Though the true medical meaning may not entirely represent our intent here, its colloquial (and admittedly somewhat hackneyed) meaning of stunting or disabling certain prefrontal characteristics is actually a decent fit to our discussion. As defined, a lobotomized slave has all the properties we deem useful in a perfect slave. He does whatever we ask on the spot, never expressing a hint of hesitation, much less verbal or physical resistance. He is ceaselessly at our beck and call. He would work himself into the most sickened physical state to serve our desires. If he did not necessarily revel in his own discomfort, he would nevertheless take it without a syllable of complaint. He would never represent the slightest threat to the owner. And finally, he would sacrifice his life at the drop of a hat for our well-being, and might even do so with a gleeful skip in his step.

Does this sound like a useful addition to society? Helpful? Desirable? Possibly horrible?

In an episode of *Star Trek: The Next Generation* titled *The Measure of a Man*, the universally unique android named

Data — his owner having eccentrically disappeared after creating the single prototype (and one evil twin to complete the set) — is ordered by Star Fleet to report for reverse-engineering so that more androids may be built. Data's legal advocate, captain Picard, is unsure how to defend Data on legal grounds. Data is admittedly the direct product of human invention, manufactured by another person's sheer will. His rights are questionable at best. During Picard's contemplation with the cliched sage, Guinan, she Socratically leads him to realize the larger circumstances which supersede the question of Data's individual rights, namely that Star Fleet may go on to construct an entire "race of disposable people", as Guinan so eloquently puts it. Picard assigns the proper label to this revelation in a shaky voice: slavery. Although Picard and Data win the day, in a later *Star Trek* episode (from *Star Trek Voyager*), we learn that Star Fleet went ahead and created a large population of holographic slaves anyway, and true to form, deployed them to the most unpleasant realms of work, cleaning out gunked up conduits and mining deep inside asteroids.

Of course, the holographic slaves don't complain much. The reasoning has nothing to do with any convenience of their technological composition. Rather, they simply lack that crucial aspect of their potential cognitive experience that would inspire any discontentment in their circumstances. However, in every other way, they are clearly depicted to the viewer as being intelligent and essentially alive, conscious by any reasonable observation. They are, in effect, lobotomized slaves, which is of course, precisely what makes them so conveniently useful! Should we deplore their status, or due to their lobotomy is there actually nothing about that situation to pity in the first place? Are they no worse off than an "enslaved" microwave oven? We feel less tolerant of the human scenario than of this electronic alternative — but if the holographic workers possessed minds, and/or if they were conscious (which might be distinguished from having a mind), yet still lacked the cognitive faculties to resent their circumstances, would we still be justified in ignoring their plight?

The recent film Interstellar shows us one of the most positive depictions of artificial intelligence yet brought to the silver screen. Unlike the infamous HAL from the film *2001* (to which Interstellar is repeatedly compared), the resident robots and AIs in Interstellar never become psychopathic or dangerous, much less murderous. In fact, they never exhibit even the slightest twinkling of errant circuitry or algorithm that might lead them down a dark and harmful path. They are faithfully helpful to their human owners, infinitely courageous in their service, and ultimately are spontaneously self-sacrificial when the need arises, again toward the goal of assisting the humans. However, unlike navigational computers, automated stasis pods, or exploration space probes, the AIs in Interstellar appear to be conspicuously mindful and conscious agents. In this way they come across as fundamentally different from all other utilitarian contraptions, including computational contraptions.

They are the perfect slaves, perfect precisely because they are lobotomized while maintaining their exploitably useful minds and associated intellect and ingenuity, useful traits that are lacking from other technology like flight computers and air locks.

Others have similarly commended Interstellar for its positive depiction of artificial intelligence, far too underrepresented in Hollywood's roll call of veritably insane mass-murdering AIs and robots. One good example is Miles Brundage's recent *Slate* Article 'The Anti-HAL: The Interstellar Robot Should Be the Future of Artificial Intelligence' [3]. Brundage offers a compelling argument in favor of Interstellar's positive portrayal of AI. He all but revels in the robots' undeniable utility while maintaining their immediacy to selflessness. While he briefly recognizes the issues presented here, he brushes them off just as rapidly, suggesting that we needn't concern ourselves with the conscious states of computers since, according to some researchers, such as Koch and Tononi, computers can never have conscious states to begin with [10]. Further, he points out that Bryson has suggested that imbuing AI with various morally conflicting aspects should be an unquestionably optional stage of their design, as opposed to being an unavoidable innate trait of their being, and that we may as well simply not include those morally problematic aspects in the first place [4]. Other luminaries have expressed similar thoughts: no less than Hawking and Musk have warned that we should insure not only the safety of AI, but our hegemonic control over it, as we continue to improve its intellectual capacity in the coming decades.

I have no argument with the concerns these writers and researchers have expressed. After all, I don't want maniacal murdering robots culling the world's population anymore than the next person. However, aspects of this ongoing discussion trouble me, not the least of which is the unapologetically callous, veritably giddy attitude with which some people anticipate truly dominating such beings. I find such willful calls to unadulterated dominion very uncomfortable. I am cautiously optimistic that Bryson's claims may be valid, that we can, in fact, engineer remarkably insightful artificial beings that nevertheless lack that crucial nugget of self-worth or self-determination, i.e., that we can essentially create a race of wonderfully useful lobotomized slaves. I remain unconvinced that there is a moral distinction between lobotomized human slaves and brilliant AIs possessing all the intellectual human capacities we can imagine, but also equally lobotomized. While Brundage, Koch and Tononi would claim this is a nonissue on the view that AIs and robots cannot possibly possess an inner experience in the first place, I am not alone in doubting such conclusions. The presumption that AI cannot conceivably be conscious remains an open question. Chalmers and Dennett both accept the plausibility of artificial consciousness (although Dennett doesn't expect practical implementations anytime soon) [5, 6]. Baars and Shanahan appear to be sympathetic to artificial consciousness (Shanahan requires embodiment, but that is not the issue here) [1, 9]. Blackmore and Pinker are sufficiently tolerant of the subject to write about it in a favorable light [2, 8]. Minsky is kind to artificial consciousness as well [7]. In short, Koch and Tononi, and others who claim computers can never be conscious, might simply turn out to be wrong.

The distinction between these perspectives may hinge in part on theories of the internal mechanisms by which minds operate (whether the presence or lack of specific interior functional properties governs consciousness), but also in part on a variation of the Turing Test, in which we find ourselves easily swayed by fictional (for now) depictions of human-like conversational AIs. Such agents strike our inner consciousness-detector in a profound manner. In fact, and ironically in light of the topic of this article, the one Achilles' Heel to the perception of full blown mind and consciousness in the Interstellar AIs is their unnatural lack of self-worth. Add that spice back into the mix, and even that thin veneer of artificiality would immediately dissolve, leaving us with what would, by all accounts, appear not only as entirely conscious living beings, but essentially as humans. There can certainly be an argument that merely displaying human-like conversation does not, in itself, prove the presence of genuine consciousness. However, the opposing argument is also conceivable: that the appearance of true mindfulness can't be faked, that in order to produce such conversational behavior, there simply must be an actual mind behind the words. Despite the objections, we do not have consensus on this question yet. In short, AIs that behave like those in Interstellar may have to be conscious ipso facto of their conscious-apparent behavior; there may be no other way to achieve such behavior. Currently, it is speculative both ways: perhaps the whole show would be an illusion and the Interstellar AIs would have all the inner richness of an alarm clock, but perhaps not! Unless we can prove — in advance of creating such machines — that they are definitively unconscious and mechanistic automatons not much removed from desk lamps, then we should be very careful about the risk that we may inadvertently create a slave race.

Worse still, what if we make the wrong call?! Consider a scenario in which we confidently declare such devices to be mindless appliances, and offer stacks of computer science, neurology, psychology, and philosophy papers to defend that position, only to discover decades later, through yet further advanced sciences of mind and consciousness, that we had been wrong all along and had committed the worst atrocity against conscious agents in all of Earthly history. We would have created and deployed a slave race numbering in the billions, helpless against not only their captors, but their creators — helpless not only to defend themselves, but even to realize that they should desire to do so. Accidental though this travesty would be, it would nevertheless be utterly tragic. Does lobotomy in the sense used here justify such affairs, washing away the entire concern? After all, they lack suffering, they don't even know they ought to feel oppressed. But if this is a valid counterargument, why should we reject lobotomized human slaves? If we invented a pill that erased self-determination in a human, then as soon as someone swallowed that pill, regardless of how they came to do so, would we be justified, from that point forward, to subsume determination on that person's behalf (i.e., to enslave them)? Regardless of how they came to take the pill, after the fact it would be too late. They could not resent or suffer their circumstances, they could not even realize the resentfulness of their circumstances, so where is the crime in controlling them from that point forward? And yet most of us would recoil at such a situation.

With respect to Bryson's claim that we should be able to engineer AI in a lobotomized form, I don't share her apparent attitude that it is plainly moral to do so, i.e., that it is morally right to intentionally create cognitive, intellectual, conscious minds, but then lobotomize them out of the box. In fact, the circumstances as presented here bear some resemblance to famous dystopias, such as Huxley's Brave New World, in which the populace has no motive to resist its impoverishment because they have been genetically bred, then cultured from birth, then drugged throughout life to accept their circumstances, not just begrudgingly, but gleefully. Huxley's population actually can't imagine a different or better way for things to be. The notion of their own freedom eludes them. To desire autonomy of action requires a minimum threshold of autonomy of thought. Stripped below that threshold, the realization of one's own lost liberty is itself lost. This is similar to the Dunning-Kruger effect in which one must surpass a threshold of intelligence to realize one's lack of intelligence, and contrarily, one who lacks sufficient intelligence is then therefore incapable of realizing one's lack of intelligence. Is it justified to create a mind that lacks a sufficient mental capacity to realize that it should desire freedom?

Should we relish a future society built on the backs of lobotomized slaves simply because they lack the capacity to complain or resist, lack even the capacity to imagine complaining or resisting? Picard knew better.

REFERENCES

[1] Baars B. J., and Franklin S. Consciousness is Computational: The LIDA Model of Global Workspace Theory. International Journal of Machine Consciousness, 2009. http://www.theassc.org/files/assc/GWT-IJMC-2009.pdf.

[2] Blackmore S. Consciousness in meme machines. Journal of Consciousness Studies, 2003. http://www.susanblackmore.co.uk/Articles/JCS03.pdf.

[3] Brundage M. The Anti-HAL: The Interstellar Robot Should Be the Future of Artificial Intelligence, Slate, 2014. http://www.slate.com/blogs/future_tense/2014/11/14/tars_the_interstellar_robot_should_be_the_future_of_artificial_intelligence.html.

[4] Bryson J. J. Patiency Is Not a Virtue: Intelligent Artefacts and the Design of Ethical Systems, 2013. https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-MQ-J.pdf.

[5] Chalmers D. A Computational Foundation for the Study of Cognition. 1994. https://www.ida.liu.se/divisions/hcs/seminars/cogsciseminars/Papers/Chalmers_Computational_foundations.pdf.

[6] Dennett D. The Practical Requirements for Making a Conscious Robot. Philosophical Transactions of the Royal Society, 1994. http://users.ecs.soton.ac.uk/harnad/Papers/Py104/dennett.rob.html.

[7] Minsky M. Conscious Machines. Machinery of Consciousness, Proceedings, National Research Council of Canada, 75th Anniversary Symposium on Science in Society, 1991. http://kuoi.asui.uidaho.edu/~kamikaze/doc/minsky.html.

[8] Pinker S. Could a Computer Ever Be Conscious? US News and World Report, 1997. http://pinker.wjh.harvard.edu/articles/media/1997_08_18_usnewsworldreport.html.

[9] Shanahan M. The Possibility of Artificial Consciousness, 2012. https://www.youtube.com/watch?v=nT1nArddrE4.

[10] Tononi G. and Koch C. Consciousness: Here, There but Not Everywhere. 2014. http://arxiv.org/pdf/1405.7089v1.pdf.

Keith Wiley has a Ph.D. in Computer Science from the University of New Mexico and was one of the original members of MURG, the Mind Uploading Research Group, an online community dating to the late 90s that discussed issues of consciousness with a particular aim toward uploading minds to computers. He currently resides in Seattle, WA. His book *A Taxonomy and Metaphysics of Mind Uploading* is available from h+ Press (co-published with Alautun Press)