Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

24(2) – September 2014

Film Review: *Transcendence*

Seth D. Baum
Global Catastrophic Risk Institute

seth@gcrinstitute.org

Journal of Evolution and Technology - Vol. 24 Issue 2 – September 2014 - pgs 79-84

Is it possible to create an artificial mind? Can a human or other biological mind be uploaded into computer hardware? Should these sorts of artificial intelligences be created, and under what circumstances? Would the AIs make the world better off? These and other deep but timely questions are raised by the recent film *Transcendence* (dir. Wally Pfister, 2014). In this review, I will discuss some of the questions raised by the film and show their importance to real-world decision making about AI and other risks. Reader, be warned: This review gives away much of the film's plot, though it also suggests ideas to keep in mind when watching.

Transcendence centers around AI researchers Evelyn and Will Caster. Will is fatally shot by a member of an extremist group, Revolutionary Independence From Technology (RIFT), as part of several simultaneous attacks on AI researchers. Before Will dies, Evelyn uploads his mind into a computer system. She then connects Upload Will to the internet, setting off a chain of events in which Upload Will, with Evelyn's assistance, takes a dominant position on Earth. (I'll use the name "Upload Will" to distinguish the upload from the pre-upload biological Will.) Evelyn eventually becomes disaffected, and, with assistance from RIFT, the United States government, and some of Evelyn's colleagues (a coalition of former enemies), persuades Upload Will to accept a computer virus designed to destroy him. Both die, and the rest of the world is forced to live on without an internet.

Compared to other AI takeover films, *Transcendence* is remarkable for its sparing use of action sequences and special effects. More attention is paid to plot, character development, and intellectual concepts. It is refreshing for a film not to hide behind eye candy. Even *Transcendence*'s penultimate fight to terminate Upload Will is largely calm, with only a few light bombs shot toward Upload Will's supporting infrastructure. The main fight is an intimate conversation between Upload Will and Evelyn as she convinces him to accept the virus – an emotional scene culminating in the two lovers dying in each other's arms, reminiscent of Romeo and Juliet. Indeed, the film is more a tragic romance than a typical war-of-the-worlds robot takeover. (One could cynically interpret it as a love story between two upper-middle-class white

Americans who care more about each other than they do about the rest of the world, and indeed are willing to risk the whole world for a chance at keeping their love alive.)

But above all, *Transcendence* is a cautionary tale about the perils of AI – or the perils of the lack of AI, depending on one's perspective. Here are some of the questions it raises, in rough order of their appearance.

What institutions should manage the development of AI? Will and Evelyn are enthusiastic about the promise of AI, but they also reject government involvement, even refusing public funding. By contrast, their colleague, Joseph Tagger, sets up his group at a national lab. Will goes so far as to disdain commercial spinoffs, preferring to focus purely on the work itself.

Similar divergence on what institutions are best for AI development can be found among actual AI experts. In a recent study, my colleagues and I asked a group of AI experts whether harmful AI was most likely if it were developed by the U.S. military, by a private company, or in an open source project. The experts tended to think that either military or open source was most dangerous, though they were split on which was the more risky. Open source creates a more transparent environment, but the military has more experience handling dangerous technology (Baum et al. 2011).

Is it good to assassinate researchers whose research is dangerous? It is chilling even to entertain such a question as anything other than fiction, but it is nonetheless important. At the beginning of *Transcendence*, RIFT is portrayed as an organization of terrorists, with the FBI seeking to bring them to justice. But by the end of the film, the FBI agent tasked with the mission comes to adopt RIFT's views on AI and collaborates with RIFT to help terminate Upload Will. (The agent does also indicate that RIFT would be a good entity to blame for anything that might go wrong along the way.) RIFT's views are ultimately vindicated, but only because Upload Will becomes highly undesirable. Had Upload Will been desirable, neutral, or at worst a minor nuisance, then RIFT would have retained its terrorist label. This suggests that it is good to assassinate researchers, but only if their research actually is dangerous. What, however, if one is unsure about the dangers? And what if assassinating them means that no one will ever know for sure if it was dangerous or not? Should researcher assassins be protected, as long as they made a good faith attempt to prevent danger? These are weighty questions.

It must be said: Ideally, assassinations and other violent tactics would not be necessary. Ideally, any risks could be explained to the scientists, their sponsors, the government, the public, and other relevant parties, so that they could be reduced peacefully and by consensus. Some of RIFT's members began as students of Will, Evelyn, and other leading AI researchers. They may not have helped their cause by abandoning their studies in favor of being a secretive and violent organization. In our own world, there are established, nonviolent structures in place for debating and regulating risky research. We should all hope that these structures are up to the task of preventing the dangers that AI and other emerging technologies pose.

That said, assassination is an option under sufficiently desperate circumstances. A close real-world analog to RIFT would be the assassination of nuclear scientists with the intent of halting threatening research programs. While the details are often unknown, some nuclear scientist assassinations are believed to have been performed by rival governments acting out of desperation (Tobey 2012). This raises questions of what qualifies as a just war in an era of high-tech weaponry (Meisels 2014). RIFT, on the other hand, is depicted as a nonstate actor, with less legitimacy in its use of force. While the real-world assassinations of nuclear scientists seek to protect the assassin's country from annihilation, RIFT's actions in *Transcendence* are aimed at protecting all of humanity, a less partisan project. Regardless of their legitimacy, nonstate actors are on the rise in the actual world, in part because of new technologies that enable smaller and more decentralized groups to accomplish more.

Are rogue nonstate actors effective at reducing risk from dangerous technologies? In other words, given the need to shut down certain research, who is best placed to do it? In the film, RIFT might have benefited from state collaboration, as its actions repeatedly backfire. Had they not shot Will (and not failed to kill him immediately), he would not have been uploaded. Then, just after his uploading, they break into Evelyn and Will's home, attempting to terminate Upload Will by force. Unfortunately, this prompts Evelyn to hurry Will onto the internet, where he can gain global influence. Later, they post a video online of a man that Upload Will has greatly enhanced (he can easily lift an 800-pound object), hoping to provoke widespread condemnation of Upload Will; instead, they motivate some new volunteers to sign up for similar physical enhancement. In each case, professional law enforcement likely could have done better – if only they'd been on board with the cause.

Is society today neglecting or otherwise underestimating the risks from AI and other emerging technologies? Within the world of *Transcendence*, RIFT would not have needed to resort to rogue assassinations if there'd been broader societal concern about AI risk, including from law enforcement agencies. Later in the film, broad support does emerge for terminating Upload Will.

A similar kind of newfound concern can be seen in the real-world psychology of risk. Risks of events that have never happened before (such as a disaster from a new technology or an unprecedented environmental change) are often underestimated, because people struggle to imagine these events happening (Weber 2006). However, when people can imagine an unprecedented event, they often overestimate the risk, out of fear of the unfamiliar (Stern 2002-2003). The film captures this latter phenomenon in its repeated use of the line "People fear what they don't understand," used to denigrate those who object to pursuing with AI and other advanced technology. And so society today will tend to underestimate the risks from AI and other emerging technologies, unless the technologies capture people's imaginations. Films like *Transcendence* can play a role in this. They should seek to convey an accurate understanding of the risks.

Under what circumstances should an AI be launched? In the film, the bullet in Will's body forces a hastened decision, eliminating the chance for careful reflection on whether it was the right one to take. Ideally, decisions to launch an AI or any other potentially transformative technology should come only after extensive reflection, safety testing, and anything else that can help ensure that the launch ends well. Unfortunately, there might not always be time for all of this. A variety of desperate circumstances could hasten decision making. This could include competition from other AI groups, especially if AI has a strong first-mover advantage in which the first AI to launch will become dominant. Another desperate circumstance would be imminent global catastrophe. For example, geoengineering is a risky technology proposed for the increasingly desperate circumstances of climate change (Gardiner 2013). Launching an AI could be another such desperate measure, if done in the hope that the AI might solve the climate change problem. (Indeed, at the beginning of the film, Evelyn hails AI as a solution to social and environmental problems, and by the end, Upload Will is regrowing forests and cleaning waterways.) In sufficiently desperate circumstances, launching radical new technologies might be a risk worth taking. Likewise, avoiding these desperate circumstances can help reduce risk by reducing the likelihood that risky technologies will be launched, or launched in haste (Baum 2014).

It is important to understand that the decision to launch an AI could be irreversible. In the film, Evelyn and other characters eventually change their minds, deciding that Upload Will should be terminated. However, the only reason they succeed in terminating Upload Will is that he decides for himself to accept the virus that will destroy him. In other words, Upload Will is in complete control of the outcome. Furthermore, Upload Will is less powerful than the AIs that many commentators believe could be created in the real-world. In technical terms, Upload Will is a fairly soft takeoff. Finally, the opportunity to terminate a real-world AI may never exist. This points to a major worry with AI and other global catastrophic risks, and in contrast with smaller risks. Because the catastrophe could be both global and permanent, there may be no second chance, no opportunity to learn from experience. This means that society must get it right the first time, every time.

Can an AI be sentient? Can we know if it is? The film repeatedly asks whether Upload Will or other AIs are sentient. It becomes a bit of a running joke: An AI will be asked if it can prove that it is self-aware, and it will reply "That is a difficult question. Can you prove that you are self-aware?" This bit of humor speaks to a very serious question about real-world uploads and AIs in general. At this time, it is unknown whether AIs can be sentient, but it is an active topic of discussion (e.g. Chalmers 2010). The question is particularly important because the sentience of AI speaks to the desirability of AI, to the (significant) extent that sentience itself is considered desirable.

Should Upload Will have been terminated? In my opinion, this is the most interesting question raised by the film. By the end, the film's characters are in broad consensus in favor of termination, concerned that Upload Will cannot be contained. But is this the right decision? The downside is that terminating Upload Will requires shutting down the internet (and possibly all other electronics; this is left unclear). The film shows the internet gone for years after Upload Will's termination; potentially it will be gone forever. Here are the key factors I see in the decision:

(1) The internet and other electronics are likely required for human civilization to achieve a desirable AI, colonize space, and other high-value outcomes.

(2) Without the internet and other electronics, human civilization can remain on Earth for no more than about one billion years.

(3) With the internet and other electronics, human civilization can colonize space and continue to exist for many orders of magnitude longer than one billion years.

(4) Upload Will was itself an AI and was moving in the direction of colonizing space and achieving other high-value outcomes.

(5) While we do not know if Upload Will was sentient, he appears in the film to be both sentient and happy.

(6) Upload Will is shown causing changes to the world that we would tend to consider good. He makes people healthier and stronger, and at least as happy; he also cleans up the environment.

(7) While we do not know if Upload Will would have continued to change the world for the better, given his initial track record it seems more likely than not that he would.

Given these factors, here is what I would say.

If the loss of internet and other electronics *would* be permanent, then Upload Will should not have been terminated. Terminating Upload Will guarantees that human civilization will never expand beyond its current state, a loss of astronomical proportions. A world with Upload Will is more likely than not to have net positive value, either through his own sentience, the sentience of others, or whatever else might have value (ecosystems, etc.). Given the net positive value, the chance that Upload Will would help achieve astronomical amounts of good outweighs whatever value an electronics-free humanity could achieve in the remainder of its time on Earth. And so, based on what the film shows, I believe its characters make the wrong decision in terminating Upload Will.

However, *if* the loss of internet and other electronics *would not* be permanent, then Upload Will should indeed have been terminated. Terminating Upload Will could buy humanity the time to think through AI more carefully, helping it make an AI launch decision with more confidence. This assumes that humanity would go on to make a better-informed launch decision. It probably would: after its experience with Upload Will it could vividly imagine AI and thus would tend to be more attentive to it.

The analysis is similar for real-world decisions about transformative technologies and global catastrophic risks (Beckstead 2013; Bostrom 2003; Maher and Baum 2013; Tonn 2007). It has even been proposed that humanity would be better off enduring a few smaller catastrophes (like Upload Will) to focus its attention on the bigger ones (Wells 2009). These truly are crucial considerations for humanity, both because so much is at stake and because we may have only one chance to get it right. By drawing attention to these considerations, and by being an enjoyable film in its own right, *Transcendence* is a good film to watch and think about.

Acknowledgments

I thank Melissa Thomas-Baum for insightful discussion during the development of this manuscript and Russell Blackford for helping revise it.

References

Baum, S. D. 2009. Film review: *District 9*. *Journal of Evolution and Technology* 20(2): 86-89.

Baum, S. D. 2014. The great downside dilemma for risky emerging technologies. Manuscript in progress for *Physica Scripta*.

Baum, S. D., B. Goertzel, and T. G. Goertzel. 2011. How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change* 78(1): 185-195.

Beckstead, N. 2013. *On the overwhelming importance of shaping the far future*. Doctoral Dissertation. Department of Philosophy, Rutgers University.

Bostrom, N. 2003. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas* 15(3): 308-314.

Chalmers, D. J. 2010. The Singularity: A philosophical analysis. *Journal of Consciousness Studies* 17(9-10): 7-65.

Maher, T. M. Jr. and S. D. Baum. 2013. Adaptation to and recovery from global catastrophe. *Sustainability* 5(4): 1461-1479.

Meisels, T. 2014. Assassination: Targeting nuclear scientists. *Law and Philosophy* 33(2): 207-234.

Stern, J. 2002-2003. Dreaded risks and the control of biological weapons. *International Security* 27(3): 89-123.

Tobey, W. 2013. Nuclear scientists as assassination targets. *Bulletin of the Atomic Scientists* 68(1): 61-69.

Tonn, B. E. 2007. Futures sustainability. *Futures* 39: 1097-1116.

Weber, E. U. 2006. Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change* 77: 103-120.

Wells, W. 2009. *Apocalypse when? Calculating how long the human race will survive*. Chichester, UK: Springer-Praxis.