Values and acausal reasoning of whole brain emulations

Complementary notes on multiverse-wide superrationality

Caspar Oesterheld

An important step in the development of humanity and potentially other civilizations in the multiverse with significant ramifications for values could be whole brain emulation, i. e. scanning a biological brain, uploading it into a computer, and then simulating its behavior at a level of detail that preserves the brain's functionality. To me, it seems likely that whole brain emulation will be possible at some point, but it is plausible that other transitions (like *de novo* artificial intelligence, i. e. artificial intelligent systems not modeled after humans or other animals) will happen earlier[1], thereby precluding an era of whole brain emulations. Following Robin Hanson, who has written extensively on the topic, I will call the computer programs resulting from this process *ems*. Since (by assumption) they are functionally equivalent to human brains, ems can do most of the things that humans can do. They can learn, work, form relationships, invent things, walk around in robotic bodies, etc. While the lack of a biological body does place some limitations on what ems can do – they cannot, for instance, have regular children – many of the sensations associated with physical activities (the taste of food, sex, etc.) can presumably be simulated. They can, however, also do many things beyond the limits of human capabilities. Being software, ems can be copied an indefinite number of times. By moving an em to a faster computer (or giving them more CPU time on their current one), one can manipulate their thinking speed. They can also travel at incredible speed and low cost via the Internet and other digital communication networks.

The advantages that ems have over biological humans plausibly make it impossible for humans to compete with them. Ems do not get sick, nor do they require payment for the costs of food, health insurance, and so on. Moreover, if one scans a few elite (i. e. smart, diligent, reliable, educated, hard-working, etc.) humans and creates millions of copies of the resulting ems, then everyone but the most competent humans will be unable to compete with them. The em economy will also be able to grow much faster than its human counterpart by virtue of not being bounded by labor at all. Once you have one em and it is cheap to build its required hardware, you can easily build more of them. Producing more human workers, on the other hand, takes a lot of time and resources. The consequences of an em-dominated society (as well as the arguments for it) are detailed in Robin Hanson's book The Age of Em 2016. For shorter introductions to the topic, see:

[1] Discussion of whether *de novo* artificial intelligence or whole brain emulation will arrive first is given in chapter 4, subsection "Artificial Intelligence" in Hanson's *The Age of Em*, as well as section V in Scott Alexander's review of the book. Of course, the two topics of AI and whole brain emulation timelines can also be discussed separately. Sandberg and Bostrom discuss whole brain emulation timelines 2008; for overviews on AI timelines, see Luke Muehlhauser's What Do We Know about AI Timelines?, and a meta-survey at AI Impacts.

- Section III of Scott Alexander's review of Hanson's book;
- Robin Hanson's TEDx talk;
- the book's website, which contains short summaries of all chapters.

We will only discuss the two issues which seem most relevant to multiverse-wide superrationality: Will ems take superrationality seriously and what will their values be?

Ems and decision theory

Starting with decision theory, I see good reasons to assume that ems will take non-causal decision theory more seriously than humans do. Whereas Newcomb's problem and prisoner's dilemmas with replicas are pure thought experiments for humans – some even argue that they cannot be set up at all (Binmore, 2007) – they can actually be implemented with ems (Yudkowsky, 2010). For example, one could take one em, create 19 copies of her, and then have them play a donation game. Although such thought experiments are unlikely to be part of the day-to-day lives of ems, reasoning about correlations between copies may become much more useful in practice, too. Whereas twins are rare and often differ significantly in by their mid-twenties in terms of experiences, most ems will have many copies (Hanson 2016, p. 155) including several recent ones (ibid., pp. 169ff.). Correlations between copies who until recently shared all experiences will be especially strong. Because copies can trust each other more, it seems plausible that they will interact with each other a lot, potentially trying to coordinate as "clans" (ibid., pp. 227f.). Overall, this means that near-copies will often interact with each other, and it thus seems plausible that they would quickly learn to use non-causal decision theories to their advantage.

While we should expect correlated decision making to be more relevant for ems than for humans, many of the arguments I outline in section "Superrational cooperation on Earth" of Multiverse-wide Cooperation via Correlated Decision Making nevertheless dampen the importance of em superrationality, possibly to the extent that it may not be very important in practice after all.

Em era values

We now turn to the values of ems. The following list is expanded from the section on values from a summary of *The Age of Em* I published in my blog. I should note that I have some reservations about Hanson's predictions in this area, because I tentatively expect higher regulation than Hanson does. Needless to say (and as Hanson acknowledges), making predictions about em-era values on Earth is already very difficult, and attempting to transfer them to other civilizations makes this even harder. Hence, these arguments should be interpreted as shifting probabilities by single percentage points at most.

- Ems have no reasons to farm animals for food or use them for testing drugs. Cognitive dissonance theory suggests that this will make ems care about animals more than humans do.
- As opposed to most humans, em copies will mostly be created on demand. That is: if you are an em, you apply for jobs (or employers offer them to you) and for every job that you get, you create a copy that fills that particular job. (In some unregulated dystopian scenarios it is also possible that ems cannot veto on whether they want to

have a copy made of themselves.) This means that the question of "will this specific life be worth living?" will be more salient to ems than humans, who can rarely predict what their children's lives will be like. They will also feel more responsible for having made the decision to live their lives (given that the decision was made by a copy), so they are less likely to resent their creation (ibid, p. 120).[2] Also, there is strong selection pressure favoring ems who consider, say, a life without much leisure to be positive (p. 123). Overall, there are selection pressures towards ems wanting to make many copies of themselves.

- There is a strong selection pressure against ems who are not willing to create *short-lived* copies of themselves. If competition is strong enough (and human nature sufficiently flexible), ems will probably still value that at least one of their copies will survive, but they would probably not disvalue the death of individual copies that much. This could lead to a moral view wherein copy clans, rather than individuals, count as the morally relevant entities. This would be similar to how many people care about protecting species rather than (and often at the cost of) preserving individuals.
- Hanson argues that ems will probably not suffer much (p. 153, 371), because their virtual reality (and even their own brain) can be so easily controlled. Given that experiencing suffering *probably correlates with caring about suffering*, this could mean that ems will care less about the suffering of others.
- Assuming that individual aspects of ems can be tweaked, they could be made especially thoughtful, friendly, and so on (p. 150).
- Because of higher competition, ems will work more (e. g. see pp. 167ff., 207) and be paid less. Hence, they will not have the resources for altruistic activities that modern elites currently have.
- People who are more productive tend to be married, intelligent, extroverted, conscientious and non-neurotic. Smarter people are more cooperative, patient, rational and law-abiding, and also tend to favor trading with foreigners. Because ems will be selected for productivity, they will tend to have these traits as well (p. 163).
  - It is somewhat unclear whether ems will be more or less religious. Apparently religious people are more productive, but they are also less innovative (p. 276, 311). Hanson expects that religions will be able to adapt to the em world (p. 312).
- Workaholics tend to be male and males are also more competitive, so the em society may consist mostly of males (p. 167).
- Due to the possibility of creating a lot of copies when an em reaches a particular age, only to destroy most of the copies later, most ems will likely be at the peak productivity age of around 40 to 50 or older (p. 202ff.). 50-year-olds tend to be less supportive of war than younger people (p. 250). Also, "older people tend to associate happiness more with peacefulness, as opposed to excitement." (p. 205)
- Most ems will not have children (p. 211f.), which could, among other things, make them more compassionate towards others (Gilead and Liberman, 2014).

[2] It is, however, still possible. Incidentally, the critically acclaimed science-fiction novel Permutation City by Greg Egan starts with a newly created copy resenting its existence.

- At some point, it may become attractive to scan children in order to turn them into ems, since they can then adapt more easily to the em world (p. 212). This could give an advantage to ruthless countries and children of psychopathic parents, who are themselves more likely to be psychopathic (Viding and Larsson, 2010; Waldman and Rhee, 2006; Farrington, 2006).
- Space will lose some appeal, because it takes ages of subjective time to travel there (p. 225).
- If male ems are "castrated" (however that would exactly work for ems) because of the gender imbalances and the obsoleteness of sexual reproduction, then they will tend to be more sympathetic (p. 285).
- "Ems can travel more cheaply to virtual nature parks, and need have little fear that killing nature will somehow kill them." (p. 303) If we assume the latter to be a main reason for humans' care for the environment (Woodcock, 2000, ch. 4.IV, section "Ecosystems are Inherently Valuable?"), we should expect ems to care significantly less about preserving nature.
- The classic targets of charity – alms, schools and hospitals – may all be less necessary in an em society (p. 302). This may lead ems to support other kinds of charity.
- "New em copies and their teams are typically created in response to new job opportunities. Such teams typically end or retire when these jobs are completed. Thus ems are likely to identify strongly with their particular jobs; their jobs are literally their reason for existing." (p. 306, also see p. 328) Maybe this implies that ems will be less involved in pursuing ethical causes to give their life a meaning.
- Ems are far more likely to be anti-substratist than humans (obviously).
- Ems may find it more natural to view consciousness as a matter of degrees rather than absolutes, seeing as em minds will differ in speed and some may exist only as partial minds (p. 341ff.).
- "[Because ems will be poorer than citizens of industrialized societies, they] seem likely to return to conservative (farmer) cultural values, relative to liberal (forager) cultural values. [...] Today, liberals tend to be more open-minded, creative, curious, and novelty seeking, while conservatives tend to be more orderly, conventional, and organized. If, relative to us, ems prefer farmer-like values to forager-like values, then ems more value things such as self-sacrifice, self-control, religion, patriotism, marriage, politeness, material possessions, and hard work, and less value self-expression, self-direction, tolerance, pleasure, nature, novelty, travel, art, music, stories, and political participation. [...] If ems are indeed more farmer-like, they tend to envy less, and to more accept authority and hierarchy, including hereditary elites and ranking by gender, age, and class. They are more comfortable with war, discipline, bragging, and material inequalities, and push less for sharing and redistribution. They are less bothered by violence and domination toward the historical targets of such conflicts, including foreigners, children, slaves, animals, and nature. [...] Farmer-like ems have a stronger sense of honor and shame, enforce more conformity and social rules, and care more for cleanliness and order (Stern et al. 2014)." (pp. 326-328.)
- "As ems have near subsistence (although hardly miserable) income levels, and as

wealth levels seem to cause cultural changes, we should expect em culture values to be more like those of poor nations today. As Eastern cultures grow faster today, and as they may be more common in denser areas, em values may be more likely to be like those of Eastern nations today." Citing Inglehart and Welzel (2010) and Schwartz et al. (2012), Hanson goes on: "Together, these suggest that em cultures tend to value technology, money, hard work, and state intervention. They may also suggest that em culture values achievement, determination, thrift, authority, good and evil, and local job protection." (p. 322f.)

- The em world will be dominated by a few (i. e. something like one thousand) copy clans, copied from humans who will tend to be selected for their eliteness. "Today, Jews comprise a disproportionate fraction of extreme elites such as billionaires, and winners of prizes such as the Pulitzer, Oscar, and Nobel prizes (Forbes 2013). (I have sought but failed to find work identifying other elite ethnicities.) This weakly suggests that Jews are also disproportionately represented among ems."

References

Binmore, Ken (2007). *Game Theory: A Very Short Introduction*. OUP Oxford.

Farrington, David P (2006). "Family background and psychopathy". In: *Handbook of psychopathy*, pp. 229–250.

Gilead, Michael and Nira Liberman (2014). "We take care of our own: caregiving salience increases out-group bias in response to out-group threat". In: *Psychol. Sci.* 25.7, pp. 1380–1387.

Hanson, Robin (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

Sandberg, Anders and Nick Bostrom (2008). *Whole Brain Emulation. A Roadmap*. Tech. rep. 2008-3. Future of Humanity Institute.

Viding, Essi and Henrik Larsson (2010). "Genetics of child and adolescent psychopathy". In: *Handbook of child and adolescent psychopathy*, pp. 113–134.

Waldman, Irwin D and Soo Hyun Rhee (2006). "Genetic and environmental influences on psychopathy and antisocial behavior". In: *Handbook of psychopathy*, pp. 205–228.

Yudkowsky, Eliezer (2010). "Timeless decision theory". In: *The Singularity Institute, San Francisco*.