MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Whole Brain Emulation and the Evolution of Superorganisms

Carl Shulman
MIRI Visiting Fellow

Abstract

Many scientists expect the eventual development of intelligent software programs capable of closely emulating human brains, to the point of substituting for human labor in almost every economic niche. As software, such emulations could be cheaply copied, with copies subsequently diverging and interacting with their copy-relatives. This paper examines a set of evolutionary pressures on interaction between related emulations, pressures favoring the emergence of *superorganisms*, groups of emulations ready to self-sacrifice in service of the superorganism. We argue that the increased capacities and internal coordination of such superorganisms could pose increased risks of overriding human values, but also could facilitate the solution of global coordination problems.

Shulman, Carl. 2010. *Whole Brain Emulation and the Evolution of Superorganism*.
The Singularity Institute, San Francisco, CA.

The Machine Intelligence Research Institute was previously known as the Singularity Institute.

1. Introduction

The field of computational neuroscience studies brain function in terms of the information processing properties of the brain, frequently constructing software models of particular features. Taken to its limits, this practice could eventually result in human *whole brain emulations*, software models that can reliably mimic the behavior of human brains at various levels of abstraction. Regardless of whether we consider such systems to possess mental states, emulations with sufficient functional similarity could substitute for humans in almost any cognitive task. Experts in the area recently released a roadmap analyzing plausible computational, scanning, and other demands to create such emulations (Sandberg and Bostrom 2008). The roadmap's estimates suggest that this might be feasible by mid-century.

Because such emulations could be freely copied and run at increased speeds, they might quickly outnumber humans and be capable of performing almost any task more cheaply. Standard economic models suggest this could produce tremendous economic growth, perhaps doubling the size of economies every few weeks or less, but also driving wages for most jobs below human subsistence level (Hanson, forthcoming). Many have suggested that such rapidly replicating and evolving minds could cause human extinction if not carefully controlled (Bostrom 2002; Yudkowsky 2008; Posner 2004; Friedman 2008; McAuliffe 2001; Joy 2000; Moravec 1999). In light of the potential impacts of emulations, a clearer picture of the factors influencing the behavior of such emulations seems valuable. Here we consider one particular factor, the evolutionary pressure for emulations to form *superorganisms*, groups of related emulations ready to individually sacrifice themselves in pursuit of the shared aims of the superorganism.

2. Advantages of Superorganisms

The evolution of kin altruism, multicellular life, social insect superorganisms, or brain emulation superorganisms depends on the benefits of cooperation. The larger those benefits, the stronger the evolutionary pressures involved. An initial survey indicates that the benefits of a willingness to self-sacrifice would be extremely high for human brain emulations. Specifically, a superorganism of such entitites could realize a much higher level of economic productivity than narrowly self-concerned individuals, and could coordinate activities where formal legal methods of coordination are unavailable, e.g., protection of intellectual property against piracy and political action.

Many of the productivity advantages stem from the ability to copy and delete emulations freely, without objections from the individual emulations being deleted. One simple way to exploit this ability to increase productivity draws on the variability of

worker productivity over time. Emulations could have their state saved to storage regularly, so that the state of peak productivity could be identified. The gap between this peak productivity and average productivity could be very large: human productivity varies dramatically depending on fatigue, recent distractions, or boredom with repetitive tasks, and no work can be done while sleeping (Van Dongen et al. 2003; Henning et al. 1997; Barger et al. 2006). With stored emulations, whenever a short task arises, a copy of the peak state emulation could be made to perform the task and immediately be deleted, so that computational power could be reallocated to a fresh copy at peak for the next job. If a job required more time than the peak state would last, e.g., to learn task-relevant information and apply it, new "snapshots" could be taken after acquiring that information and used to make copies to perform short subtasks of the overall effort. This procedure might multiply emulation productivity severalfold for any task that can be done quickly (e.g., in under an hour), but at the cost of the deletion of enormous numbers of short-lived emulations. Members of a superorganism would willingly sacrifice themselves to be replaced by another member, where self-concerned individuals would prefer to escape.

The productivity benefits mentioned above might act like a one-time flat multiplier of output, allowing more emulation work to be done with a given amount of computational power. However, even more important benefits could lie in an enhanced ability to cumulatively improve the "human capital" of emulations. Educational efforts to boost emulation productivity could be vastly improved through controlled experimentation: subject thousands or millions of copies of an emulation to varying educational techniques, test their resulting performance, and use tools or emulations that have performed best to build the template for the next "generation" of emulations, deleting the rest to free up computational resources. Similarly, the brain emulation software could be altered to mimic the effects of drugs, neurosurgery, genetic changes, and other interventions. Experiments with such alterations would likely render emulations cognitively impaired or mentally ill in most cases, but in some cases might result in enhanced productivity. The ability to delete failed experiments and reallocate computational power to new ones would be essential to make such explorations feasible. Some of the resulting educational techniques and software changes might be specific to idiosyncrasies of the experimental subjects, while other methods, once identified, could be applied to unrelated emulations. The more such methods are idiosyncratic, the greater the cumulative advantage for superorganisms.

The methods outlined above to enhance productivity could also be used to produce emulations with trusted motivations. A saved version of an emulation would have particular motives, loyalties, and dispositions which would be initially shared by any copies made from it. Such copies could be subjected to exhaustive psychological testing, staged

situations, and direct observation of their emulation software to form clear pictures of their loyalties. Ordinarily, one might fear that copies of emulations would subsequently change their values in response to differing experiences (Hanson et al. 2007). But members of a superorganism could consent to deletion after a limited time to preempt any such value divergence. Any number of copies with stable identical motivations could thus be produced, and could coordinate to solve collective action problems even in the absence of overarching legal constraints.

3. Evolutionary Routes to Superorganisms

We have seen that superorganisms would enjoy some advantages, but how easy would it be for superorganisms to develop in the first place? And what mechanisms would translate those advantages into reproductive success?

We have defined emulation superorganisms in terms of the willingness of the individual members to sacrifice themselves in pursuit of the shared aims of the superorganism. The specific basis of this willingness is not essential: many different combinations of values and beliefs might generate the relevant behavior, including combinations that appear to exist in some individuals today. Views on personal identity and survival vary, but many individuals who have considered the question agree with Derek Parfit that, instead of personal identity, what matters is "Relation R: psychological connectedness and/or continuity, with the right cause" (perhaps *any* cause) (Parfit 1986, 215), where "psychological connectedness" means "the holding of particular direct psychological connections" and "psychological continuity" means "the holding of overlapping chains of *strong* connectedness" (Parfit 1986, 206).

A recent survey asked philosophers about Parfit's Teletransporter case, in which an individual's original body is destroyed, and a functionally identical copy is constructed elsewhere (similar to the process of "moving" a computer file from one medium to another). More than a third of target faculty accepted or leaned toward the view that the individual survives this process, while slightly less accepted or leaned toward the view that the individual dies (PhilPapers 2009). This question only considers future "descendants" of an individual, and not "copy-siblings" who were copied from a common "ancestor," but it does suggest a willingness to adopt expansive accounts of personal survival and identity. Views that emphasize psychological similarity might treat deletion of one emulation, to be replaced by a copy made the night before, as no worse than waking up after a night of carousing without memory of the event (Hanson 1994).

Some might instead prize a narrow concept personal identity but hold values such that accepting deletion to uphold those values would nevertheless be acceptable. A soldier willing to sacrifice her life for her country might be willing to do so a hundred

million times, provided that other copies are available to take up her life projects and commitments. A consequentialist might endorse switching one emulation for another that will realize more good. More likely, the best candidates for emulation would reflect some mixture of labor productivity, views about personal identity, and values.

After emulations are created, further psychological tactics could be deployed to strengthen these motivations, using the educational experimentation discussed above. Interestingly, even an individual who initially cared only about her own future selves, i.e., her "copy-descendants" and not her "copy-siblings," would wish to use such methods to change her descendants: if A creates B and C, and cares equally about them, A's preferences are likely to be best satisfied if B and C care equally about one another.

Once brain emulations are cheap enough to substitute for human labor, a competitive market might resemble a hybrid of modern software markets and Malthusian population growth. If initially the price of rented computer hardware or cloud computing resources was less than the wages that skilled brain emulations could earn, more copies could be made until the price of computing resources was bid up to equal plummeting wages (Hanson 1994, 1998b, forthcoming). In perfectly competitive markets, this would result in wages for work done by brain emulations just sufficient for emulation subsistence, i.e., rented computer hardware, bandwidth, etc. At that intensity of competition, even a modest productivity advantage for a new lineage of emulations could allow it to outbid competing emulations for resources, rendering the necessities of existence unaffordable for self-concerned emulations dependent on wages. The productivity advantages discussed earlier could easily lead to superorganisms of one sort or another overwhelmingly dominating an unregulated market, and thus making up a majority of the emulation population.

The intellectual property created by experimentation with emulation education and enhancement would also create a niche for profits by superorganisms. To fund costly experimentation, experimenters must capture some revenue: if an enhanced emulation simply reproduced as much as it could, it would rapidly drive wages down to subsistence, capturing relatively little of the surplus value generated by the improvement. Like software companies, those improving emulation capabilities would need methods to prevent unlimited unlicensed copying of their creations. Patents and copyrights could be helpful in this respect, using the legal system to punish unlicensed reproduction, but the ethical and practical difficulties would be great. In particular, the incentives for piracy would be unprecedently great since emulation labor would be a much greater fraction of business costs than any software programs today. However, a superorganism, with shared stable values, could refrain from excess reproduction and capture maximum profits (to fund further experimentation or other projects) without drawing on the legal system for enforcement.

The market considerations discussed above might be circumvented by regulation (although enforcement might be difficult without emulation police officers, perhaps superorganisms for value stability) in a given national jurisdiction, but such regulations could impose large economic costs that would affect international competition. With economic doubling times of perhaps weeks, a major productivity or growth advantage from self-sacrificing software intelligences could quickly give a single nation a preponderance of economic and military power if other jurisdictions lacked or prohibited such (Hanson 1998b, forthcoming). Other nations might abandon their regulations to avoid this outcome, or the influence of the less regulated nation might spread its values as it increased its capabilities, including by military means. A sufficiently large economic or technological lead could enable a leading power to disarm others, but in addition a society dominated by superorganisms could also be much more willing to risk massive casualties to attain its objectives.

Consider a contest between two powers, one populated by superorganisms and one by self-concerned individuals. Each possesses weapons of mass destruction, such that a preemptive strike would completely destroy the other power, although retaliatory action would destroy 90% of the inhabitants of the aggressor. For the self-concerned individuals, this would be a disaster: each would face a 90% chance of death, a terrible evil by their lights. But for the superorganisms, the loss of 90% of their members, which could be rapidly replaced as new hardware is constructed using the spoils of war, would be no worse than the normal deletion and replacement of everyday affairs.

The combination of competitive dynamics within and between regulatory jurisdictions would thus tend to result in a predominance of emulations fitting our definition of superorganisms, barring some scheme of global governance capable of controlling evolutionary pressures, i.e., a "singleton" (Bostrom 2006).

4. Implications for Existential Risk

Nick Bostrom defines an existential risk as "[o]ne where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom 2002). Bostrom has also identified two major classes of existential risks posed by human brain emulation. First, he raises the possibility of a global takeover by human brain emulations with objectionable values as an instance of drastically curtailed potential, specifically describing a scenario in which the brain emulation discovers methods to enhance its own intelligence, which it uses to further enhance its intelligence, and so forth, and imposes objectionable values on our future. Second, he warns that unrestrained evolutionary competition among human brain emulations could result in human values such as happiness, conscious experience, and humor largely ex-

punged from our future (Bostrom 2004). The features of emulation superorganisms discussed above suggest that emulations enable the formation of a singleton more easily than might otherwise be expected. This increases the first risk, the danger of a singleton with objectionable values, but reduces risks of unrestrained competition in the absence of a singleton.

The key feature of superorganisms in this context is their ability to produce many copies with matching stable values. These saved states could be copied billions of times to staff an ideologically uniform military, bureaucracy, and police force. After a short period of work, each copy would be replaced by a fresh copy of the same saved state, preventing ideological drift. Within a given jurisdiction, this capability could allow incredibly detailed observation and regulation: there might be one such copy for every other resident. This could be used to prohibit the development of weapons of mass destruction, to enforce regulations on brain emulation experimentation or reproduction, to enforce a liberal democratic constitution, or to create an appalling and permanent totalitarianism, itself an existential risk (Caplan 2008).

Further, this superorganism ability could also be used to enforce agreements between superorganisms. Consider two mutually distrustful nation-states, each of which has a governmental apparatus that is composed of copies from a distinct superorganism. A treaty to restrict dangerous technology development between the two would founder without an inspection and enforcement mechanism. But treaty enforcement/inspection activities could be conducted by joint teams of copies of each member lineage in the ruling coalition. A monitoring team could be embedded in an open-source program protecting the digital intelligences within from tampering, and all members could jointly inspect sensory feeds to the system. Such a monitoring team would observe and test every member of the global citizenry for treaty violations, with independent mechanisms available to each team member to alert the broader world to defections. Communications channels could be restricted so that messages would be short (preventing codes) and public. If monitoring teams were also regularly reverted to saved states, this could allow every coalition member to allow the infiltration of its territory by coalition monitoring teams without fear that the information gained by the monitors might be used to attack the monitored country (save in punishment for a treaty violation). Scaling up these protocols could integrate numerous superorganisms to produce a global singleton.

Why bother with such extraordinary coordination techniques? Among other reasons, to avoid the existential risks of unrestrained Malthusian competition among emulations. In a Malthusian environment, where any nontrivial increase in productivity would allow a superorganism to expand enormously, while falling behind could push it below subsistence level, there would be enormous pressure to push the bounds of experimentation with emulation software. If some alterations enhance productivity in exchange for in-

creased risk of insanity or change of values, competitive pressures could force widespread adoption of the risky changes (Yudkowsky 2008). With time this could result in a population that previous generations of humans and emulations would have considered utterly devoid of value (Bostrom 2004). In the course of interstellar colonization, selection on competing colonizers could cause them to "burn the cosmic commons," evolving to expend almost all available resources on faster colonization, diverting most accessible resources to this purpose rather than to generating lives worth living or other goods (Hanson 1998a). This would be a particularly tragic outcome because of the vast amounts of such goods that could be created by a more orderly colonization (Bostrom 2003).

Thus, it seems plausible that if emulations could solve the coordination problems of creating a singleton (before dispersal in space rendered this infeasible), they would attempt to do so. However, in the time preceding success, evolutionary pressures might result in an emulation population with quite inhuman values, which humanity would consider a loss of intelligent life's potential. We might do best by attempting to shape the character of early emulations in benevolent directions, through careful selection, testing, and cultivation. Those systems could then form superorganisms and a global singleton with relatively greater human influence and less divergence of emulation and human values.

This has been a very brief sketch of some ideas about one aspect of emulation evolution and behavior. However, if we take the possibility of successful human brain emulations this century seriously, and find some plausibility to the above discussion, more rigorous and extensive analysis would be well worth the effort.

References

Barger, Laura K., Najib T. Ayas, Brian E. Cade, John W. Cronin, Bernard Rosner, Frank E. Speizer, and Charles A. Czeisler. 2006. "Impact of Extended-Duration Shifts on Medical Errors, Adverse Events, and Attentional Failures." *PLoS Medicine* 3 (12): e487. doi:10.1371/journal.pmed.0030487.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. http://www.jetpress.org/volume9/risks.html.

———. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing,* edited by Charles Tandy, 339–371. Vol. 2. Death and Anti-Death. Palo Alto, CA: Ria University Press.

———. 2006. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48–54.

Bostrom, Nick, and Milan M. Ćirković, eds. 2008. *Global Catastrophic Risks.* New York: Oxford University Press.

Caplan, Bryan. 2008. "The Totalitarian Threat." In Bostrom and Ćirković 2008, 504–519.

Friedman, David D. 2008. *Future Imperfect: Technology and Freedom in an Uncertain World.* New York: Cambridge University Press.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

———. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." Unpublished manuscript, July 1. Accessed April 26, 2012. http://hanson.gmu.edu/filluniv.pdf.

———. 1998b. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. http://hanson.gmu.edu/longgrow.pdf.

———. Forthcoming. "Economic Growth Given Machine Intelligence." *Journal of Artificial Intelligence Research.* Preprint at. http://hanson.gmu.edu/aigrow.pdf.

Hanson, Robin, James Hughes, Michael LaTorra, David Brin, and Giulio Prisco. 2007. "The Hanson-Hughes Debate on 'The Crack of a Future Dawn.'" *Journal of Evolution and Technology* 16 (1): 99–126. http://jetpress.org/v16/hanson.pdf.

Henning, Robert A., Pierre Jacques, George V. Kissel, Anne B. Sullivan, and Sabina M. Alteras-Webb. 1997. "Frequent Short Rest Breaks from Computer Work: Effects on Productivity and Well-Being at Two Field Sites." *Ergonomics* 40 (1): 78–91. doi:10.1080/001401397188396.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. http://www.wired.com/wired/archive/8.04/joy.html.

McAuliffe, Wendy. 2001. "Hawking Warns of AI World Takeover." *ZDNet UK Edition,* September 3. http://www.zdnet.com/hawking-warns-of-ai-world-takeover-3002094424/.

Moravec, Hans P. 1999. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

Parfit, Derek. 1986. *Reasons and Persons.* New York: Oxford University Press. doi:10.1093/019824908X.001.0001.

PhilPapers. 2009. "The PhilPapers Surveys: Preliminary Survey Results." November. Accessed April 27, 2012. http://philpapers.org/surveys/results.pl.

Posner, Richard A. 2004. *Catastrophe: Risk and Response.* New York: Oxford University Press.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Van Dongen, Hans P. A., Greg Maislin, Janet M. Mullington, and David F. Dinges. 2003. "The Cumulative Cost of Additional Wakefulness: Dose-Response Effects on Neurobehavioral Functions and Sleep Physiology From Chronic Sleep Restriction and Total Sleep Deprivation." *Sleep* 26 (2): 117–126. http://www.journalsleep.org/ViewAbstract.aspx?pid=25803.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008, 308–345.