MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Singularity Summit 2011 Workshop Report

Anna Salamon, Luke Muehlhauser
Machine Intelligence Research Institute

Salamon, Anna, and Luke Muehlhauser. 2012. *Singularity Summit 2011 Workshop Report*. Technical Report, 2012-1. The Singularity Institute, San Francisco, CA.

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

Workshop Participants

Anna Salamon (organizer)	Researcher, Singularity Institute
Luke Muehlhauser (organizer)	Researcher, Singularity Institute
Seth Baum	Graduate student, Pennsylvania State University
Nick Beckstead	Graduate student, Rutgers University
David Brin	Author and futurist
David Chalmers	Philosopher, Australian National University
Paul Christiano	Graduate student, MIT
Kevin Fischer	Entrepreneur
Alexander Funcke	Owner, Zelta Deta
Julia Galef	Science writer
Phil Goetz	Researcher, J. Craig Venter Institute
Katja Grace	Graduate student, Carnegie Mellon University
Robin Hanson	Economist, George Mason University
Louie Helm	Director of Development, Singularity Institute
Jey Kottalam	Entrepreneur
Nathan Labenz	Co-founder, Stik.com
Zvi Mowshowitz	Co-founder, Variance
Luke Nosek	Co-founder, Paypal
Carl Shulman	Researcher, Singularity Institute
Tim Sullivan	Co-founder, CO2Stats
Jaan Tallinn	Co-founder, Skype
Michael Vassar	President, Singularity Institute
Alexander Wissner-Gross	Research Affiliate, MIT Media Lab
Eliezer Yudkowsky	Researcher, Singularity Institute

Executive Summary

The Singularity Summit 2011 Workshop was held at the Marriott Courtyard Hotel in New York City on October 17–18, 2011. Participants discussed many topics in small and large groups. This report includes just two of the results:

1. The results of the longest single discussion, about whole brain emulation.
2. A list of questions to which participants said they would like to know the answer.

AGI Safety and Whole Brain Emulation

Should humanity accelerate WBE development?

Shulman and Salamon (2011) argued that if whole brain emulation (WBE) came before artificial general intelligence (AGI), this might help humanity to navigate the arrival of AGI with better odds of a positive outcome. The mechanism for this would be that safety-conscious human researchers could be "uploaded" onto computing hardware and, as uploads, solve problems relating to AGI safety before the arrival of AGI.

Shulman and Salamon summarize their position this way: 'Business as usual' is not a stable state. We will progress from business as usual to one of four states that *are* stable: stable totalitarianism, controlled intelligence explosion, uncontrolled intelligence explosion, or human extinction. (E.g., "Go extinct? Stay in that corner.") The question, then, is "whether the unstable state of whole brain emulation makes us more or less likely to get to the [stable] corners that we [favor]. Is our shot better or worse with whole brain emulation?"

Shulman and Salamon suggested that our chances of achieving a controlled intelligence explosion are *better* if whole brain emulation arrives before AGI. Why? Shulman and Salamon gave two major reasons.

First, WBE could ease coordination problems. If AGI is developed in an arms race, there would be strong incentives for AGI developers to trade off safety for faster development. But whole brain emulation might be safer because we would be dealing with a relatively well-understood mind architecture (the human one), and this could enable developers to design the relevant safety architectures in advance. A slight lead in development of WBE could provide a very large lead once WBE is reached, because WBEs could be run at 1000x or more the normal speed (if hardware is available). Thus, a four-year lead in developing AI could be turned into the equivalent of a 4,000-year lead if the leader develops WBE first. The leader's researchers could be uploaded and continue their work at 1000x or more the normal human speed. This would give the leader in AGI development substantial subjective time (e.g. 4,000 "years") to develop safe AGI. (Shulman and Salamon also outlined other advantages possessed by uploaded researchers.) Without WBE, a four-year lead is only a four-year lead, providing strong incentives to prefer development speed over development safety.

Second, WBE could provide powerful tools for safe control of uploaded safety researchers. One could monitor uploaded researchers and "reboot from disk" if danger to others arose. (However, this raises ethical concerns and may be morally equivalent to killing the researcher.)

If it is correct that WBE would improve humanity's chances of successfully navigating the first creation of AGI, we might be able to accelerate WBE development—e.g.

by increasing funding for computational neuroscience (Sandberg and Bostrom 2008)—to increase the odds that WBE will arrive before AGI. This would be an instance of safety-conscious "differential technological development" (Bostrom 2002).

But if pushing on WBE development instead *increases* existential risk, then we may wish to avoid accelerating that field of research.

Participants in this discussion began in broad disagreement about whether it would be wise to accelerate WBE research for the purposes of future AGI safety. They agreed, however, that Friendly AI is the safest form of AGI if it is possible, that WBE is the next-safest, that neuromorphic (neuroscience-inspired) AI is the next safest after that, and that non-brain-inspired ("de novo") AI is the least safe (apart from Friendly AI). They also agreed that if humanity accelerates WBE-related technologies, this increases both the probability of WBE and the probability of (dangerous) neuromorphic AI. Participants modeled this as if there were an "enough brainscanning" variable (with an unknown probability distribution), and another probability that turns "enough brainscanning" into either WBE or neuromorphic AI.

Participants disagreed about whether WBE or neuromorphic AI was likely to come first from pushing on WBE-related tech. Zvi Mowshowitz, who was appointed "keeper of the probabilities," integrated participants' opinions to estimate that the probability that neuromorphic AI would arrive before WBE was 0.85. (Zvi was given this role because he was relatively unbiased on the subject and had done tens of thousands of such calculations throughout his work experience.) The considerations discussed included:

1. So far, we've made rapid progress in figuring out which parts of the brain do what, and how it all works;
2. Before you can build a complete WBE you'll have emulated neurons, and these are things you can play around with to observe their functioning under varying conditions; but on the other hand:
3. Merely reverse-engineering the Microsoft Windows code base is hard, so reverse-engineering the brain is probably much harder.

When discussing the likely effects of not-provably-Friendly *de novo* AI and provably Friendly AI, participants agreed that non-Friendly AI was a far more likely outcome than Friendly AI. Intelligence amplification (Bostrom and Sandberg 2009) was unanimously considered to raise the probability of good outcomes. Nanotechnology was unanimously considered to lower the probability of good outcomes, because it was agreed that it might cause disaster and would at least cause people to take riskier actions.

After some discussion, participants gave their estimates for the probability of a good result ("win") given various scenarios: e.g. prob(win | WBE or neuromorphic AI), prob(win | WBE), etc. Individuals' estimates varied quite a bit. Zvi gave his personal

estimate from a weighted average of others' estimates, weighted by his estimate of each participant's expertise. His given estimates were a roughly 14% chance of win if WBE or neuromorphic AI comes first and, coincidentally, a roughly 14% chance of win if de novo AI (Friendly AI or not) came first.

The original plan for deciding whether to recommend the acceleration of WBE-relevant tech was to figure out which win probability was larger: prob(win | WBE or neuromorphic) or prob(win | de novo AI).

The tiebreaker consideration deemed most important was timelines. Pushing on WBE-related tech would, if anything, cause machine superintelligence to be created sooner. All else being equal, sooner timelines would seem to decrease the probability of a win scenario. Thus, despite many shortcomings in the design of this expert elicitation, the group's current collective best guess, given the considerations evaluated so far, is that humanity should not accelerate WBE-related tech.

Tough Questions

At another point during the workshops, participants wrote down a list of questions for which knowing the answer would plausibly affect their immediate actions. Below is a complete list of the questions.

- Can you measure curiosity? How do you increase it?
- Do values converge?
- How well does dual n-back work?
- Can large groups be reasonably goal-seeking instead of just doing the thing that the group does?
- How much does rationality help ventures?
- How plausible are certain AI architectures? When should they arrive?
- How do the benefits of rationality training compare to other interventions like altruism training?
- How hard is it to create Friendly AI?
- What growth rates can be sustained?
- What is the strength of feedback from neuroscience to AI rather than brain emulation?
- Is synthetic biology easier and likely to to become a threat sooner than we realize?

- How can people nominally motivated to reduce x-risk be practically motivated?
- To what degree can salesmen be replaced by bots?
- How sane is the NSA? Can they control anything?
- What is the virality and influence on a rationality community of rationality lessons/books depending on their polish?
- Is there a way to turn money into good writing?
- Is there a safe way to do uploads, where they don't turn into neuromorphic AI?
- Can we get a post-doc to write up Timeless Decision Theory?
- Is CEV misguided? How do we evaluate it?
- What effect do skeptics and Hacker News people have upon the rationality community when recruited?
- How possible is it to do FAI research on a seastead?
- How much must we spend on security when developing a Friendly AI team?
- Can we "raise the sanity waterline" much at all?
- What is the future scalability of quantum computers?
- Will molecular manufacturing ever work?
- What's the best way to recruit talent toward working on AI risks?
- How much money do I need to live in the future, what will be the rate of return on investments?
- Under which instances do heuristics win over rationality?
- How can we identify good ideas?
- How difficult is Friendly AI vs. intelligence explosion?
- How can intermediate AI code be kept secure?
- How difficult is stabilizing the world so we can work on Friendly AI slowly?
- How much goal drift will occur over time?
- How hard will a takeoff be?
- How much experimentation does a machine need to self-improve?
- How much are human goals maximizing or satisficing?

- How difficult is correct philosophical consensus?
- What is the value of strategy vs. object-level progress toward a positive Singularity?
- How much interest is there in an existential risk conference or a Friendly AI conference?
- What topology does the graph of relatively rational people have?
- Is the nature of reality completely different than we currently think?
- How feasible is Oracle AI?
- Can we convert environmentalists to people concerned with existential risk?
- Is there no such thing as bad publicity for our purposes?
- Which actions have been identified as clearly useful but are not being performed because of social norms?
- What are the payoffs from specific Singularity Institute programs?
- On which questions should I delegate to others' beliefs?
- Which skills would I get the most use from per hour of training?
- How can I buy more attention and energy?

References

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.

Bostrom, Nick, and Anders Sandberg. 2009. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311–341. doi:10.1007/s11948-009-9142-5.

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.

Shulman, Carl, and Anna Salamon. 2011. "Whole Brain Emulation as a Platform for Creating Safe AGI." Talk given at the Fourth Annual Conference on Artificial General Intelligence, Mountain View, CA, August 3–6.