Simulation and the Singularity

David Chalmers

Departments of Philosophy
Australian National University
New York University

The Intelligence Explosion

- "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

  - I.J. Good, 1965

Terminology

- AI: intelligence of human level or greater
- AI+: intelligence of greater than human level
- AI++: intelligence of far greater than human level.

Argument

- (1) There will be AI+.
- (2) If there is AI+, there will (soon afterwards) be AI++.
- ———————————
- (3) There will be AI++.

From AI to AI++

- (1) There will be AI (before long, absent defeaters).
- (2) If there is AI, there will be AI+ (soon after, absent defeaters).
- (3) If there is AI+, there will be AI++ (soon after, absent defeaters).
- ———————————————
- (4) There will be AI++ (before too long, absent defeaters).

Timeframe and Defeaters

- Before long = within centuries
  - Maybe pessimistic, but 2035 is optimistic
- Soon after = within years
- Defeaters = disaster or active prevention.

Premise 3:
From AI+ to AI++

- Assume we create an AI+.
- An AI+ will be better than us at AI creation.
- So it will be able to create a greater AI than we can.
- So it will be able to create an AI greater than itself.
- Repeat until AI++.

Premise 2:
From AI to AI+

- Assume we create an AI *by an extendible method.*
- Then we will inevitably soon improve the method, and create an AI+.
- N.B. This requires an extendible method.
  - Biological reproduction is not.
  - Nor is brain emulation.

Premise 1:
The Path to AI

- Why believe there will be AI?
- Evolution got here, dumbly.
- We can get here too.

Intelligence Measures

- Intelligence isn't unitary. But...
- G [weakly] measures capacity A if increasing G(x) [tends to] increase A(x).
- G is a self-transmission measure if G measures the capacity to create systems with G.
- G is a general intelligence measure if G weakly measures cognitive capacities A, B, C, ...

Generality Thesis

- The generality thesis: There is a self-transmitting general intelligence measure
  - G such that increasing G increases capacity to create systems with G and tends to increase cognitive capacities A, B, C, ...
- Then we can substitute G for "intelligence" in the previous arguments.

Different Paths to AI

- Direct programming: Really hard.
- Brain emulation: Not extendible.
- Learning: Maybe, but still hard.
- Simulated evolution: Where my money is.

Self-Improving Intelligence

- The intelligence explosion argument turns on humans creating human+ intelligence.
- But it works just as well if any system S can create S+ intelligence
- Even if S is much dumber than us, intelligence will explode.

Evolution as Self-Improvement

- If S is dumber than us, S can't directly create S+ intelligence.
- But S can still yield S+ intelligence, through evolution.
- Simulated evolution is in effect a dumb path to superintelligence.

Evolutionary AI

- Getting (open-ended, powerful) simulated evolution right is a hard unsolved problem.
- But (my bet) an easier problem than getting learning right, or getting intelligence right.
- Nature did it from very few resources.

AI in Simulated Worlds

- If we arrive at AI+ through simulated evolution, it will very likely take place in a simulated world.
- Not in a robot or other system directly interacting with our environment.
- If so, what follows, practically and philosophically?

Negotiating the Singularity

- Q: How can we negotiate the singularity, to maximize the chances of
  - (i) a valuable post-singularity world
  - (ii) a valuable post-singularity life for ourselves and our descendants.

Negotiating the Singularity

- I. Advance planning
- II. Ongoing control.

AI+ in a Simulated World

- AI+ in a simulated world offers us somewhat more control.
- We are not inhabiting a common environment.
- We can make initial observations about AI + and make decisions about how to proceed.

The Leakproof Singularity

- Ideal: The Leakproof Singularity.
- A "leakproof" simulated world, constructed so that laws of the simulation do not allow systems to leak out.
  - No red pills!
- The AI+/AI++ within it does not immediately leak out into our world.

The Leakproof Singularity II

- A fully leakproof singularity is impossible, or pointless.
- A non-pointless singularity can be observed
- When we observe, information leaks out.

The Leakproof Singularity III

- Leakage of systems is under our control
- If we communicate with AI+/AI++, they will soon leak out.
- If they have information about us, likewise.

The Leakproof Singularity IV

- The key to a controllable singularity is not to preventing information from leaking out, but to prevent information from leaking in.
- Autonomous simulated worlds, closed systems without any ongoing input from us.
- Design may provide some hints for AI++, so idiosyncracies of design should be minimized.

The Leakproof Singularity: Summary

- I. Create AI in simulated worlds.
- II. No red pills.
- III. No external input.

Benign and Non-Benign Worlds

- If a post-singularity simulated world is not benign, we can try again.
- If it is benign, we can attempt integration in a controlled way.

Integration into a Post-Singularity World

- Q: How do we integrate with a post-singularity simulated world?

- A: By uploading and self-enhancement.

- [Alternatives: separatism, inferiority, extinction.]

Questions

- I. Will an uploaded system be conscious?
- II. Will it be me?

Consciousness

- We don't have a clue how a computational system could be conscious.
- But we also don't have a clue how a brain could be conscious.
- No difference in principle?

Gradual Uploading

- Upload one neuron at a time, preserving organization throughout.
- Will consciousness fade or disappear?
- I've argued: it will stay constant.

Organizational Invariance

- Consciousness is an organizational invariant
- Systems with the same pattern of causal organization have the same sort of consciousness.

Personal Identity

- Will an uploaded system be me?
- Personal identity is not an organizational invariant.
- My twin and I are different people.

Three Views

- Any two systems with the same organization are the same person
  - Implausible (twins)
- Same person requires same matter
  - Implausible (neuron replacement)
- Same person requires causal connectedness
  - Yes -- but what sort?

Continuity of Consciousness

- Best sort of causal connection: continuity of consciousness.
- Gradual uploading, staying conscious throughout.
- One stream of consciousness.
- Comparable to ordinary survival?

Reconstructive Uploading

- If I'm dead (and brain is unrevivable), gradual uploading is impossible.
- But there's still reconstructive uploading, from records
  - Brain, brain scans, audio, video, books
- AI++ could reconstruct causal organization from this.

Reconstructive Identity

- Will the reconstructed system be me?
  - Pessimistic view: It's like my twin surviving
  - Optimistic view: It's like waking up.

Buddhist View

- Ordinary surviving is like my twin surviving
  - Each waking is a new dawn
  - And that's good enough
- If so, reconstructive uploading will also be good enough.

Practical Question

- Q: How can we encourage AI++ to reconstruct us?
- A: Write articles and give talks about the singularity.

The End