University of Bern
Discussion in the CAS AML M5 Philosophy and Ethics of Extended Cognition and Artificial Intelligence"
October 25th, 2024

Paper: Chalmers, David J. (2010), The singularity: A philosophical analysis, Journal of Consciousness Studies 17 (9-10), 7–65

presentation by Cem and Rolf

In what follows, we give a short overview of the paper with discussion points from our session.

1. Research questions

- will there be a singularity?
- if there will be, what are the practical implications?

2. Chalmers's method

- clarification of "singularity": here idea that there will be an intelligence explosion, i.e., superintelligent AI (i.e. intelligent systems that are way more intelligent than are humans now) due to Good's mechanism in the not too distant future. There are other ideas, e.g. that of a mathematical singularity that are not at the center of Chalmers's paper.
- The questions are highly speculative. Chalmers thus investigates arguments for there being a singularity. In this way, possible pathways, scenarios and options become clearer.

3. Discussion

- General discussion: Do you think that there will be a singularity in the sense defined?
  - The amount of energy required for a singularity might be too large.
  - Counter: AI might become smart enough to find new technologies to obtain energy.
  - What exactly does it mean that there is AI that is surpasses humans in terms of intelligence? What is part of intelligence? E.g. emotions? Maybe, AI can never feel emotions.
  - Another issue is creativity; it's arguable that human intelligence includes creativity; the question then is to what extent AI can become creative.
  - An argument against the singularity: We build our own mistakes into AI; for instance, some AI is suffering from biases.
  - Possible counter: This doesn't exclude that AI becomes in important respects superior to humans and that Good's mechanism is realized.
  - What's important for Chalmers is Good's mechanism. So, the AI has at least to be capable of building more intelligent systems. It may be that certain aspects of intelligence are not realized in AI, but then, if AI becomes significantly more intelligent than humans in some respects that include the building of more intelligent systems, this would already be quite remarkable and worthy of discussion.
  - Another argument: We don't have enough motivation to build such systems.
  - Possible counter: There are many humans; it's likely that at least some of them have an interest in building AI and AI+. Probably not that many people are needed to do so.
  - So far, AI was admittedly focused on systems that can achieve very specific tasks. But in some areas, there is a huge interest in AI that can achieve various goals, e.g. care robots should be able to do various things.

    - The question about Chalmers's AI is what is otherwise called the question about artificial general intelligence (AGI).
- How might we get to AI? Chalmers:
    - brain emulation, so far not yet fully done.
    - Mimic evolution
    - machine learning
- What are possible defeaters?
    - We discussed the role of war; might hamper technological evolution, but has often fostered it.
    - In any case, the development of AI+ depends on various contingent factors.
    - Might AI, AI+ destroy itself? That depends on its values/goals. It's plausible that AI+ will have the goal to preserve itself to some extent; it would be pointless to build systems that soon destroy themselves
- How should we evaluate, and prepare ourselves for, a singularity assuming it has a non-negligible probability?
    - Chalmers distinguishes between different evaluations; while a singularity may not be good to humans, it may be good overall in some sense, for instance, because AI systems may become more valuable than humans.
    - In any case, we have a strong interest that AI doesn't do things that destroy what we take to be valuable.
    - Objection: There are no shared humans values, so what is it that "we take to be valuable".
    - Possible counter: Despite value disagreements, we agree about the value of some things, e.g. human life. For sure, our values have changed, but we think that there has been some progress, and we want to avoid, for instance, that AI harms humans; see Asimov's laws.
    - Still, when building AI, we cannot do justice to all values that have been defended. So the builders of AI have to start with their own values. This is related to the alignment problem to which we will come.
- Scenarios for the future with AI
    - extinction
    - isolation: problem: not attractive to isolate AI from humans because it can benefit humans
    - inferiority
    - integration
- General points:
    - Is Chalmers exaggerating risks?
    - The paper is 13 years old; today, do we have more evidence against or for a singularity.

Summary: The singularity is certainly a scenario that is quite speculative. Still, we think about its likelihood. In the next few sessions, we will return to questions raised by current AI.

cb