Is ethical AI possible?

June 21, 2025

> Thinking systematically about the risks of AI requires two lines of thinking. First of all, we have to establish what it is we want and don't want. Secondly, we then have to find the most effective ways of imposing our preferences.[1]

Structure of the program

I chose the program's clickbaity title because it seems that thinking about AI within the framework of ethics is a way to get at a broad area of what one would want from AI. The title (and the similarly biased image of the Tower of Babel) also betray my own cautious approach to AI: like game-changing inventions/discoveries from the wheel and fire to nuclear power, AI has potential benefits and potential hazards. We need to think about the latter, especially when evangelists are preaching (and hyping) the former.

One of my goals for this program is to have a discussion of what it is that we (the people in this room) would want from AI. This means that today's program is more within the humanities than it is in technology (though I tried to study up on the tech to be able to answer tech questions that might come up).

But beyond today's discussion, I'm hoping that this program (and others like it) can get an ongoing discussion going. Generative AI seems to have taken us over a threshhold away from an era where AI was only useful in limited domains (e.g. expert systems for doctors) to one where AI (of a sort) is used broadly. If that's right, then AI will be affecting us all, so we might all want to think about how to manage it.

Some possible topics for discussion

- What would we like AI to be able to do for us?

[1] Susskind, *How to Think about AI*, 110.

- What do we *not* want AI to be able to do for us?
- How do we want it trained?
    - Is it reasonable for those creating an AI to feed it a lot of material that is covered by copyright? After all, we have no problem with a human reading all that material and synthesizing it into new ideas, as long as the human properly ccites sources and doesn't just repeat substantially the same material as what they read.
    - Or does the ability of AI to repeat passages as if the entire work has been memorized (even though it hasn't) indicate a qualitative difference that breaks any such analogy?
    - Or does the analogy still hold (at least somewhat) but we'd want the AI to observe the same rules as a human: if you think you might be reproducing what you read too closely, go back and do a better job of expressing the ideas in your own way?
    - Tangentially, what does the conflict between innovation and copyright say about the term of copyright, a term that went from 14 years with the possibility of 1 renewal back in 1790 to the lifetime of the author plus 70 years as it is today (to oversimplify the status of copyright term)? If the lag on material entering the Public Domain were shorter, would that resolve the conflict?
- Do we want it used in education?
    - If it's used in the workplace, we probably want kids to learn how to use it at some point in their education, but do we only want it to get introduced near graduation, just before students go out into the workplace?
    - Or is there a role for AI in enhancing education earlier? If so, how do we avoid/reduce the ***calculator effect*** of overdependence?
    - And just like thinking about AI opened a can of worms for copyright, would we need to rethink how we educate and evaluate if AI makes it trivial to game the system of evaluation?
- What will AI do to employment? to particular professions? to the economy generally?
    - Some jobs will become obsolete. Will AI use result in enough new jobs to replace them?

    - Will AI use create demand for non-AI related goods and services and, thus, non-AI related jobs?
    - Who should get the economic benefits that come from AI and any reduction in the amount of work that humans need to do?
- When should we trust AI to be accurate?
    - When I got my first calculator, I'd do calculations by hand as well as on the calculator so that I could assure myself that the calculator was trustworthy (see the comment about my cautious nature in the introductory paragraph lol). That didn't last too long since the calculator agreed with my calculations.
    - Flash-forward 18 years from then to the Pentium FDIV bug. The bug only affected division and then only if one were dividing certain pairs of numbers. If a similar bug had affected my calculator back in 1976, it's not likely that my redundantly calculating by hand would have caught it.
    - Errors in arithmetic algorithms are (relatively… *very* relatively) "easy" to catch, but if we're relying on AI to do a job, how do we know that it has come up with correct information?
    - Of course, the same question comes up if we delegate a project to a human. Can we use the methods that we have for judging the human assistant's work in judging that of the AI? What does it mean for AI to be good enough?
- When should we trust AI? Period
    - How do we feel about AI processing our personal data?
    - Does an AI have sufficient controls on what it "remembers" as it works to avoid leaking data from one domain into another?
- In gathering information, what biases will AI be operating under? Note that the list of possible biases includes the biases of society in general, the so-called ***Overton window***.
- What will be the environmental impact of AI?
    - Electricity?
    - Water (used in cooling the computers)?

  - How does that compare with other technologies that we now take for granted (e.g. the internet generally)?
- What ethical issues arise if we ever get “real” AI (***Artificial General Intelligence*** or ***AGI***)?
  - Currently, our AI is very definitely unthinking.
  - Even large language models that seem fluent in English do not have anything like an explicit grammar[2] but rely instead on statistical likelihood of word triples (***trigrams***), so it is generous even to call them “language models,” let alone “intelligent.”
  - This gets pretty blue-sky-y, but if AGI does come about, does intelligence imply consciousness?
  - Note that people are capable of acting in what might be unconscious states (e.g. when so drunk that they later have no memory of what they did while intoxicated).

Resources

Susskind, Richard E. *How to Think about AI: A Guide for the Perplexed*. 1st ed. Oxford, England ; New York, N.Y.: Oxford University Press, 2025.

Image credits

The Tower of Babel
By Peter Bruegel the Elder
License: Public Domain

[2] One could question to what extent humans construct grammatical structures in the course of their language use, but decades of research in psycholinguistics and neurophysiology seem to indicate that they do construct them.