Re111: *AI and the Gorilla Problem*

retraice.com

Russell and Norvig say it's natural to worry that AI will destroy us, and that the solution is good design that preserves our control.

Our unlucky evolutionary siblings, the gorillas; humans the next gorillas; giving up the benefits of AI; the standard model and the human compatible model; design implications of human compatibility; the difficulty of human preferences.

Air date: Monday, 9th Jan. 2023, 10:00 PM Eastern/US.

The gorilla problem

Added to Re109 notes after live:

"the **gorilla problem**: about seven million years ago, a now-extinct primate evolved, with one branch leading to gorillas and one to humans. Today, the gorillas are not too happy about the human branch; they have essentially no control over their future. If this is the result of success in creating superhuman AI—that humans cede control over their future—then perhaps we should stop work on AI, and, as a corollary, give up the benefits it might bring. This is the essence of Turing's warning: it is not obvious that we can control machines that are more intelligent than us."[1]

We might add that there are worse fates than death and zoos.

Most of the book, they say, reflects the majority of work done in AI to date—within '**the standard model**', i.e. AI systems are 'good' when they do what they're told, which is a problem because 'telling' preferences is easy to get wrong. (p. 4)

Solution: uncertainty in the purpose (**the 'human compatible' model**[2]), which has design implications (p. 34):

- chpt. 16: a machine's **incentive to allow shut-off follows from uncertainty** about the human objective;
- chpt. 18: **assistance games** are the mathematics of humans and machines working together;
- chpt. 22: **inverse reinforcement learning** is how machines can learn about human preferences by observation of their choices;
- chpt. 27: problem 1 of N, our choices depend on **preferences that are hard to invert**;
  problem 2 of N, **preferences vary** by individual and over time.

The human problem

But how do we ensure that AI engineers don't use the dangerous standard model? And if AI becomes easier and easier to use, as technology tends to do, how do we ensure that ***no one*** uses the standard model? How do we ensure that no one does ***any* particular thing**?

The 'human compatible' model indicates that the 'artificial flight' version of AI (p. 2), which is what we want, is **possible**. It does not indicate that it is **probable**. And even to make it probable would still not make the standard model improbable. Nuclear power plants don't make nuclear weapons' use less probable. This is the more general problem taken up by Bostrom (2011) and Bostrom (2019).

□

References

Bostrom, N. (2011). Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy*, *10*, 44–79. Citations are from Bostrom's website copy:
https://www.nickbostrom.com/information-hazards.pdf Retrieved 9th Sep. 2020.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. Nov. 2019. Citations are from Bostrom's website copy:
https://nickbostrom.com/papers/vulnerable.pdf Retrieved 24th Mar. 2020.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking. ISBN: 978-0525558613. Searches:
https://www.amazon.com/s?k=978-0525558613
https://www.google.com/search?q=isbn+978-0525558613
https://lccn.loc.gov/2019029688

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson, 4th ed. ISBN: 978-0134610993. Searches:
https://www.amazon.com/s?k=978-0134610993
https://www.google.com/search?q=isbn+978-0134610993
https://lccn.loc.gov/2019047498

[1] Russell & Norvig (2020) p. 33.

[2] Russell (2019).