Re115: *Technological Progress, Defined* (Technological Danger, Part 3)

retraice.com

How we would decide, given predictions, whether to risk continued technological advance.

Danger, decisions, advancing and progress; control over the environment and 'we'; complex, inconsistent and conflicting human preferences; 'coherent extrapolated volition' (CEV); divergence, winners and losers; the lesser value of humans who disagree; better and worse problems; predicting progress and observing progress; learning from predicting progress.

Air date: Friday, 13th Jan. 2023, 10:00 PM Eastern/US.

Progress, 'we' and winners

If the question is about 'danger', the answer has to be a decision about whether to proceed (advance). But how to think about progress?

Let '**advance**' mean moving forward, whether or not it's good for humanity. Let '**progress**' mean moving forward in a way that's good for humanity, by some definition of good.[1]

Progress can't be control over the environment, because *whose* control? (Who is *we*?) And we can't *all* control equally or benefit equally or prefer the same thing. This corresponds to the Russell & Norvig (2020) chpt. 27 problems of the complexity and inconsistency of human preferences,[2] and Bostrom (2014) chpt 13 problem of "locking in forever the prejudices and preconceptions of the present generation" (p. 256).

A possible solution is Yudkowsky (2004)'s 'coherent extrapolated volition'.[3] If humanity's collective 'volition' doesn't converge, this might entail that there has to be a 'winner' group in the game of humans vs. humans.

This implies the (arguably obvious) conclusion that we humans value other humans more or less depending on the beliefs and desires they hold.

Better and worse problems can be empirical

Choose between A and B:

- carcinogenic bug spray, malaria;
- lead in the water sometimes (Flint, MI), fetching pales;
- unhappy day job, no home utilities (or home).

Which do *you* prefer? This is empirical, in that we can ask people. We can't ask people in the past or the future; but we can always ask people in the present to choose between two alternative problems.

Technological progress

First, we need a definition of progress in order to make decisions. Second, we need an answer to the common retort that 'technology creates *more* problems than it solves'. 'More' doesn't matter; what matters is whether the new problems, together, are 'better' than the old problems, together.

We need to define two timeframes of 'progress' because we're going to use the definition to make decisions: one timeframe to classify a technology *before* the decision to build it, and one timeframe to classify it *after* it has been built and has had observable effects. It's the difference between *expected* progress and *observed* progress. Actual, observed progress can only be determined retrospectively.

Predicted progress:

A technology *seems like* progress if:
the predicted problems it will create are better to have than the predicted problems it will solve,
according to the humans alive at the time of prediction.[4]

Actual progress:

A technology *is* progress if:
given an interval of time, the problems it created were better to have than the problems it solved,
according to the humans alive during the interval.

(The time element is crucial: a technology will be, by definition, progress if *up to a moment in history* it never caused worse problems than it solved; but once it does cause such problems, it ceases to be progress, by definition.)

Predict*ion* progress (learning):

'Actual progress', if tracked and absorbed, could be used to improve future 'predicted progress'.

□

[1] Retraice (2022/10/24).

[2] Cf. Russell & Norvig (2020) p. 34 and Re111 (Retraice (2023/01/09)).

[3] See also Bostrom (2014) p. 259 ff.

[4] The demonstrated preferences of those humans? The CEV of them? This is hard.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford. First published in 2014. Citations are from the pbk. edition, 2016. ISBN: 978-0198739838. Searches:
https://www.amazon.com/s?k=978-0198739838
https://www.google.com/search?q=isbn+978-0198739838
https://lccn.loc.gov/2015956648

Retraice (2022/10/24). Re28: What's Good? RTFM. *retraice.com*.
https://www.retraice.com/segments/re28 Retrieved 25th Oct. 2022.

Retraice (2023/01/09). Re111: AI and the Gorilla Problem. *retraice.com*.
https://www.retraice.com/segments/re111 Retrieved 10th Jan. 2023.

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

Yudkowsky, E. (2004). Coherent extrapolated volition. *Machine Intelligence Research Institute*. 2004.
https://intelligence.org/files/CEV.pdf Retrieved 13th Jan. 2023.