The Perils and Promise of Artificial Conscientiousness

Matt Beane

We humans are notoriously bad at predicting the consequences of achieving our technological goals. Add seat belts to cars for safety, speeding and accidents can go up. Burn hydrocarbons for cheap energy, warm the planet. Give experts new technologies like surgical robots or predictive policing algorithms to enhance productivity, block apprentices from learning. Still, we're amazing at predicting unintended consequences compared to the intelligent technologies we're building.

WIRED OPINION

ABOUT

Matt Beane (@mattbeane) is an assistant professor of technology management at UC Santa Barbara and a research affiliate at MIT's Institute for the Digital Economy.

Take reinforcement learning, one particularly potent flavor of AI that's behind some of the more stupendous demonstrations as of late. RL systems take in reward states (aka goals, outcomes that they get “points” for) and go after them without regard to unintended consequences of their actions. DeepMind's AlphaGo was designed to win the board game Go, whatever it took. OpenAI's system did the same for *Defense of the Ancients* (*DOTA*), a fiendishly complex, multiplayer online war game. Both came up with unconventional, in some cases radical, new tactics required to beat the best that humanity had to offer, yet consumed disproportionately large amounts of energy and natural resources to do so. This kind of single-mindedness has inspired all kinds of fun sci-fi, including an AI designed to produce as many paperclips as possible proceeding to destroying the earth, and then the entire cosmos, in an effort to get the job done.

While seemingly innocuous, this win-at-any-cost approach is untenable with the more practical uses of AI. Otherwise we may end up swamped by power outages, flash-trading market failures, or (even more) hyper-polarized, isolated online communities. To be clear, these threats are possible only because AI is delivering amazing improvements on previous best practices: electrical grids are becoming much more efficient and reliable, microsecond-frequency trading allows for major improvements in global market efficiency, and social media platforms suggest beneficial connections to goods, services, information, and people that would otherwise remain hidden. But the more we hand these and similar processes over to AI that is singularly focused on its goals, the more they can produce consequences we don’t like, sometimes at the speed of light.

Some within the AI community are already addressing these concerns. One of the founders of DeepMind cofounded the Partnership on AI, which aims to direct “attention and effort on harnessing AI to contribute to solutions for some of humanity’s most challenging problems.” On December 4, PAI announced the release of SafeLife, a proof-of-concept reinforcement-learning model that can avoid unintended side effects of its optimization activity in a simple game.

SafeLife has a clear way of characterizing those consequences: increases in entropy (or the degree of disorder or randomness) in the game system. By definition this is not a practical system, but it does show how a reinforcement-learning-driven system can optimize towards a goal while minimizing collateral damage.

This is very exciting work, and in principle it could help with all kinds of unintended effects of intelligent technologies like AI and robots. For example, it could help factory robots know they should slow down if a red-tailed hawk flies in their way. (I've seen this happen. Those buildings house pigeons, and, if big enough, birds of prey). A SafeLife-like model could override its programmed setting to maximize throughput, because destroying living things adds a lot of entropy to the world. But some things that we expect to help in theory end up contributing to the very problems they're trying to solve. Yes, that means the unintended consequences module in next-gen AI systems could be the very thing that creates potent unintended consequences. What happens if that robot slows down for that hawk while a nearby human expects it to keep moving? Safety and productivity could be threatened.

This is particularly problematic when these consequences span significant amounts of space and time. Take the *DOTA* algorithm. During a match, when it calculates its win probability is above 90 percent, it's programmed to taunt other players via chat. "Win probability 92 percent," you might read as you watch your hard-won forces and devious strategy decimated by a computer program. What effects does that have on players' approaches to the game? And, even further removed, what about their commitment to the game? To gaming generally? Their career aspirations? Their contributions to society? If this seems like armchair speculation, note that Lee Sedol—the world's best professional Go player, a wunderkind who has devoted his entire life to mastering the game—has just quit the game publicly and permanently, saying that no human can beat the system. It's not obvious that Sedol's retirement is good or bad for the game, for him or for society, but it is a symbolic and significant unintended consequence of the actions of an AI-based system optimizing on its reward function.

This goes way beyond games. We already know that the AI underlying social media optimizes on things like click-through and time on the site, which has led us into social echo chambers where we interact only with people who share our views. The causal links between these kinds of outcomes and the actions of these algorithms are so difficult to sense, measure, and account for that it's hard to imagine any AI designer incorporating them into a SafeLife-inspired system.

We will make progress here, of course, and as with other digital technologies, the pace and extent of that progress is likely to surprise us. But there's significant risk here too. We may make progress on entropy-aware AI in a way that leads many of us to think that we've got unintended AI consequences *covered*, and we won't then pay attention to the possibility of other unintended effects that are much more complex and play out over longer timescales. This has happened with something as inert and easy to observe as seat belts, so we should expect a rougher ride with distributed IT like AI. We might guard against this with enough attention from those familiar with dynamics of complex social systems and by expanding the range of input data available to these systems, but then again, these might be the very things that ensure its occurrence. History will be the ultimate arbiter here, and we may well live to see a world filled with safer AI, self-replicating paperclips, or—more likely—a bit of both.

More Great WIRED Stories

- Why the “queen of shitty robots” renounced her crown
- Amazon, Google, Microsoft—who has the greenest cloud?
- Instagram, my daughter, and me
- Ewoks are the most tactically advanced fighting force in Star Wars
- Everything you need to know about influencers
- 👁 Will AI as a field "hit the wall" soon? Plus, the latest news on artificial intelligence
- 🏃 Want the best tools to get healthy? Check out our Gear team’s picks for the best fitness trackers, running gear (including shoes and socks), and best headphones.