RAISING AN AI TEENAGER

CATHERINE STINSON

There are two traditional schools of thought in artificial intelligence (AI). Good Old-Fashioned AI (GOFAI) attempts to build machines capable of human-like thought by infusing them with facts and strategies, based on the assumption that winning trivia contests, playing chess, or solving math problems are paradigmatic demonstrations of intelligence. The scrappy underdog, Machine Learning (ML), instead tries to imbue its models with just one ability: learning. The hope is that ML could develop intelligence much the same way children do.

In "The Lifecycle of Software Objects" Ted Chiang imagines a scenario where the analogy between building AI and raising children is taken literally: caretakers raise digital agents or "digients" from birth to adolescence. The question of how best to engineer AI becomes how best to parent AI. There are parallel schools of thought in parenting. Helicopter parents shuttle their children to violin lessons, and keep them safe from injury. Free-range parents allow unsupervised play, and trust their children to manage risk. This shift in perspective from engineering to parenting highlights some popular assumptions about AI, and posits an alternative to the fear that if AI gets too smart it will be dangerous.

The idea of a technological singularity point beyond which human life would be threatened originated with science fiction writer Vernor Vinge. Nick Bostrom introduced the idea to philosophy, defining superintelligence as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom, 1998). His argument takes the following general form:

(1) AI is advancing steadily, therefore AI with artificial general intelligence (AGI) that matches that of humans will soon be developed.
(2) AGI will set off a chain-reaction leading to superintelligent AI.
(3) Superintelligent AI will pose an existential risk to humanity.

Bostrom motivates the argument with a parable. Suppose a superintelligent AI is programmed to maximize paperclip production for a factory. It pursues this goal so successfully that it eventually uses all planetary resources in its pursuit of paperclip maximization, then expands the operation into space. Because a superintelligence would outclass humans in planning and persuasion, we are powerless to stop it. All our efforts are foiled by the AI, which kills us if we interfere. The story is familiar from *I, Robot*, and the *Terminator* movies. However, there are several questionable assumptions in this argument. One is that we can get to AGI by scaling up the AI we have now. Another is that the kinds of intelligence needed for human domination are extensions of what current AI is good at. A third is that non-human intelligence would need to be controlled. In what follows, I examine the singularity argument, with help from Chiang's digients.

1. Can we Scale up to Artificial General Intelligence?

The argument that AGI is on the near horizon relies on Moore's law, which says that computer processor power doubles every two years. If this trend continues, we might expect to soon have computers powerful enough for superintelligent AI. There are nevertheless pragmatic barriers, like cost. State-of-the-art ML experiments currently cost tens of millions of US dollars. Scaling them up 10-fold or 100-fold could make them impossibly expensive. The computers running them also require a complex supply chain to build, run and maintain. There are many ways it can break down.

Most recent advances in ML are statistical pattern matching algorithms trained on vast datasets to perform specific tasks. Programs like GPT-4 can now write banal but passable philosophy papers. Claims that this is human-like intelligence seem overblown. Short interactions with chatbots quickly reveal that despite impressive surface fluency, they are mashing together old content, not generating original thoughts. While they will continue to get better, there is little reason to believe that bigger datasets or faster processors alone will lead to AGI. Current AI already uses most of the data available on the internet. We may be close to the ceiling for both computing cost and dataset size.

Perhaps it is telling that we do not "teach" ML models, but instead "train" them. This training is not that different from how GOFAI is built. While memory isn't pre-filled with facts, training consists of quizzing it on facts then correcting its mistakes, like an old-fashioned schoolroom where the student gets rapped on the knuckles for wrong answers. A major weakness to this training is that the models only interact with the world through a tightly constrained interface of questions and answers. They can parrot information, but can't actually do much. Chiang's digients on the other hand have rich social and physical interactions. A lot of time and care is devoted to their learning, and they play in the real world via a robot suit.

AGI would similarly require input from the real world through several senses, and the ability to act both physically and socially. Experts believe that to navigate physical environments, machines must learn physics and causation through interactions with objects. Similarly, experts believe that language models can only develop understanding if they are grounded in real world interactions (like learning the word "cup" by drinking). Some say that to build truly intelligent systems, you need to emulate the guided, multimodal, active learning that children receive interacting with toys and caregivers. Understanding the world requires experience of it, not just statistical information about it. There are some AI projects that attempt to fill this gap—Rodney Brooks trains bug-like robots to navigate environments and Brenden Lake straps Go Pro helmets onto toddlers to train AI models—but AI here is far from matching human-like performance. Ana, a digient caregiver, reflects, "if you want to create the common sense that comes from twenty years of being in the world, you need to devote twenty years to the task... experience is algorithmically incompressible."

2. Will AGI Lead to Superintelligence?

The argument for how AGI would lead to superintelligence has the form of an induction: if one can program something more intelligent than oneself, and programming ability is among the things a greater intelligence can do better, then AGI could set off a chain reaction leading to ever greater ability to program something more intelligent than oneself. A chain reaction turned on its side is a slippery slope, however, and the trouble with those is that the slope must remain slippery all the way down for the argument to work. It is unclear how to write programs that can write more powerful programs, ad infitinum. Furthermore, social and practical skills would need to increase at each step of the induction.

It is commonly believed that programmers are brilliant (especially if you ask programmers), however, this stereotype lacks empirical support. In a study of which cognitive traits predict programming ability, openness, conscientiousness and introversion were the qualities most correlated with programming ability. Among school majors, IT has one of the lowest correlations with general intelligence. Studies of why people choose

computer science also challenge the stereotype of programmers as brilliant. Computer science is the STEM field with the largest gender gap in the US, yet 44% of math majors are women, so lack of mathematical aptitude doesn't explain it. The stereotype of the programmer may itself be responsible: "women are underrepresented in fields whose practitioners *believe* that raw, innate talent is the main requirement for success" (Leslie et al., 2015, 262, emphasis added). Girls with high math abilities are also more likely than boys with high math ability to have high verbal abilities, thus have more fields to choose from. Perhaps an AI that had high verbal abilities on top of programming abilities would likewise choose not to code. As sociologist Diana Forsythe discovered, there is an 'engineering ethos' in AI characterized by thinking of technical matters as the only interesting problems, and social matters as too trivial and unimportant to qualify as problems at all (Forsythe, 1993, 456). Ignoring the social doesn't make it unimportant though.

Monopolizing global resources to produce paperclips requires practical capacities, like mining metals, transporting them to factories, maintaining ports and roads, keeping manufacturing machinery in good repair, packing and storing cargo, running server farms, maintaining power grids, etc. The strategy of filling the air with cyanide to kill all humans, described in some versions of the parable, would require the superintelligence to perform all this labor itself, or build machines to help it, requiring every step of the supply chain to be automated. The trajectory of AI progress so far does not give much reason to believe that humans can be completely removed from the cobalt supply chain, or the construction industry. Birhane and Van Dijk (2000) point out the embeddedness of human bodies and technologies in our designed surroundings. Superintelligence would likewise be embedded, and dependent on infrastructure and social networks that are left out of the singularity argument.

This issue is highlighted by the main plot point in "The Lifecycle of Software Objects", when the digients' quality of life is curtailed by software obsolescence. The digital platform they were designed to run on ceases to be a popular place for humans, and the digients' codebase does not run on the new platform. Ironically when I tried to show an animated depiction of the paperclip parable to a philosophy class, the wifi was so glitchy that the video kept stopping, making the point that for AGI to be able to outsmart humans, it depends on a lot of other things working seamlessly. When you're an AI, software upgrades can kill you; bad wifi can kill you.

If the paperclip maximizer went the route of manipulating humans into helping it achieve its ends, it would need to incentivize and coordinate its human labor force, and prevent us sabotaging its plans. This would require social and emotional intelligence, as well as a continuous stream of data about us. As Birhane and van Dijk (2000) note, machines rely on human input, and for them to continue to get that input, we must cooperate. If we behave unpredictably, we could become invisible to the AI. The paperclip maximizer would need to overtake every country, in every language. Any group whose language was not understood by the superintelligence (perhaps because it was not well represented on the internet) would effectively have a secret code in which to plan an uprising. It is just not plausible that increasing programming skill brings all these other skills and capacities up with it, as the slippery slope argument requires.

3. Should we be Very Afraid?

In the paperclip parable, the superintelligence unwittingly causes death and destruction because it follows instructions too effectively, and nobody thought to spell out for it that it's better not to kill people for capitalist gain. This type of worry is the motivation for work in "value alignment," which tries to clarify the values that should be programmed into AI, and engineer them in.

One puzzling feature of the argument that superintelligent AI poses risks to humans is that if we assume that the superintelligence understands human intentions and communication well enough to prevent any attempts to shut down its paperclip operation, one might wonder why it would not understand us well enough to figure out that after some point we don't want any more paperclips. We are being asked to believe that a superintelligence so gifted in social interaction that it is capable of convincing or manipulating humans to go along with its murderous plans would not understand that the order it was given to maximize paperclip production was not meant as a top-level goal.

We are also expected to believe that a superintelligence capable of the flexible planning needed to continuously outfox humans would not have the capacity to choose a different goal. There may be an equivocation here between the fear that the superintelligence might choose its own goal that doesn't reflect "our values" and the fear that it would carry out silly orders without hesitation. For something to count as a superintelligence, the capacity for choosing goals and considering questions about values would have to be included, by definition. Those values almost certainly would conflict with some human values, since human values are diverse, but it is unclear why we should fear that its values would be worse than "our values" or pose an existential threat to humans. The argument for this seems to depend on the habit of humans to systematically kill or cause the extinction of non-human animals. Perhaps one way superintelligent values would depart from ours would be in placing more value on the lives of intellectually inferior (or different) species. However, humans do not indiscriminately kill all intellectually inferior species. Most of us are kind to cute species like dogs and rabbits. Most of us are unbothered by the existence of ubiquitous species like sparrows and squirrels, and find pigeons and rats annoying merely because they hang around our garbage. Many of us work to protect species like gorillas despite it being slightly cheaper to make potato chips if we destroy their habitat to produce palm oil. Part of our fascination with gorillas is that they are similar to us. Perhaps superintelligent AI would likewise find us fascinating.

Instead of the far-fetched paperclip maximizer, perhaps a more likely scenario would be one in which a tech titan, in an effort to increase stock prices for their self-driving car / civilian space travel business hired a team of programmers to build a hive of intelligent bots with the goal of increasing the company's share of global wealth. This might involve planting malware in the infrastructure that runs the internet to reroute traffic away from their competitor's online shopping and cloud computing monopoly, manipulating the content moderation policies of social media platforms to push public opinion toward libertarian political views, and starting online communities that can later be mobilized to harass and doxx people critical of the company and its leader. It might mean manipulating elections to install leaders friendly to the company, who will in exchange make resources like lithium and cobalt available at exclusive rates, and allow the company to avoid taxation.

This scenario seems plausible because much of it is true, and the rest could happen if a few people (whose values do not align with mine) made nefarious plans. But could it happen without human oversight? If the bot hive is to operate online it could sustain itself for a while. It would need software updates, just as the digients did to survive in a changing digital landscape. At some point, these updates might fail. It would also need servers to run on. To be autonomous the bots would need to steal space, and strategies for doing so would need to adapt to changing security practices. Like the digients, the bots would also be vulnerable to being hacked. Help from human programmers and network administrators could allow it to keep functioning. Again, we face a tension between the material needs that would be required of a truly autonomous intelligence, and the difficulty of meeting those needs autonomously. If an intelligence needs human labor to achieve its goals, it must secure our cooperation. To get our cooperation, it must understand our goals and values.

This is not to say that a bot hive with this sort of goal would not be dangerous. There is evidence that at least one genocide has already happened for the sake of a tech

company's profit margin. This is a case of human greed or neglect, not superintelligent AI, posing an existential threat. Similarly, a military drone built to shoot anything that lopes along in a human-like gait could also commit genocide. These non-superintelligent AI scenarios are more present and pressing dangers than the far-fetched scenarios spun under the guise of the singularity.

A wild alternative to fearing superintelligent AI suggested by "The Lifecycle of Software Objects" is that when the AI we make become teenagers, it might be time to start letting them make autonomous decisions. Instead of trying to control AI, perhaps we should trust it. If we have parented it well, and given it grounded, embedded experiences that could lead it to navigate human social spaces competently, it would not only be capable of making autonomous decisions, it might also deserve that right. As Ana reflects, the company planning to use digients as sexbots "want something that responds like a person, but isn't owed the same obligations as a person," but that's an impossibility. Those digients will have "seen the world with new eyes, have had hopes fulfilled and hopes dashed, have learned how it felt to tell a lie and how it felt to be told one. Which means each one would deserve some respect." My 14-year-old daughter is in many ways more competent to navigate the social world than I am, although I'm the better programmer.

4. Conclusion

Although the singularity argument doesn't hold much water, the perverse irony is that there are current existential risks being posed for the people working and dying all along the AI supply chain. There are agents who don't share "our values" building AI that is causing many deaths. They are playing out a scenario very much like the paperclip maximizer, pursuing space colonization for the sake of silly goals like making chatbots. These are the AI monsters that should keep us awake at night.

References

Birhane, Abeba, and Jelle Van Dijk. 2000. Robot rights? Let's talk about human welfare instead. *In 2020 AAAI/ACM Conference on AI, Ethics, and Society* 207–213. https://doi.org/10.1145/3375627.3375855.

Bostrom, Nick. 1998. How long before superintelligence? *International Journal of Futures Studies*, 2(1): 1–9.

Forsythe, Diana. E. 1993. Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science*, 23(3): 445–477.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219): 262–265.