What Are You Afraid Of? AI Doesn't Kill People; People Kill People

Roger Schank[1], Ray Bareiss[1]

[1]Socratics Arts

Back in the late 60's, I worked at Stanford AI Lab which had plenty of money from the Defense Department, and I was trying to figure out what to build. I was trying to get computers to understand problems and computers to understand English. They had a big robotics group. One day, they were asked to build killer robots – this was the 60s at the height of the Vietnam war. Stanford was full of anti-war people who basically refused to do it. These days there is still plenty of funding for AI, but for the most part neither the government nor business knows what AI is.

The things everyone says about AI are stupid - hard to believe. I recently saw something by Elon Musk's girlfriend talking about AI bringing communism to the world. We heard how Steven Hawking was afraid of AI. Every other day an article in newspapers or online says something frightening about AI. What about AI are they frightened of? I've worked in AI in 50 years, and the world still is not sure what AI is. I know what I was working on; I wanted to understand how people understand and to build machines to understand - not for evil reasons. I'm tired of people talking to Siri and believing it understands what they are saying. It doesn't. It would be good for the world to have a machine that could really understand us and actually communicate meaningfully.

People worry about machines spying on us. Is that a legitimate concern? What if the machine seems to understand what you are saying, but it really doesn't? A more legitimate concern is people believing that machines understand and making (bad) decisions accordingly. Most business and government leaders have no idea how an AI system makes decisions and how reliable those decisions are, but they trust them anyway.

I mentioned killer robots. Back then we didn't know how to build killer robots, and no one working in AI wanted to build them them anyway. Well, the Defense Department did want killer robots, and trust me, they still do (now called autonomous drones). Instead of worrying about AI, we need to worry about intrusive government and increasingly about intrusive businesses. The point is that AI, per se, is not evil, but it can be used (or more likely misused) with evil consequences.

I'm sitting on a comfortable chair right now. I could take that chair, turn it over, swat someone, and kill them. Should we not build chairs for that reason? These questions about evil of AI are stupid, but they keep appearing. The Defense Department wants to use AI to make more

AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA

roger@socraticarts.com (R. Schank); ray.bareiss@socraticarts.com (R. Bareiss)
http://www.socraticarts.com/ (R. Schank); http://www.socraticarts.com/ (R. Bareiss)

effective weapons. Of course they do; that's what they do. Don't blame this on AI since no one really knows what AI is; blame it on the people who want to misuse this technology.

We saw the same sorts of problems in the 80's with expert systems. The idea was to interview an expert to capture his or her rules about thinking in that domain, to program the rules into computer, and the computer would then be an expert. But computers didn't know what they were doing, and outside of a narrowly constrained domain they were disastrously incompetent. But the media went crazy; there were news shows about expert systems – how they would replace people and how they were evil. The fact that they never worked very well was lost. As we get better at building machines with intelligence, they will work better. They won't be intrinsically evil, but they will be capable of evil because of the people who use them. I maintain that the real issue is not about controlling AI; rather the issues are human ignorance and bad intentions (which have existed as long as people have existed).

1. Unintentional Evil

Let's set aside the issue of AI systems purposely engineered to do evil and deal with the more insidious problem of AI systems doing unintentional "evil" because of sloppy or naïve implementations. This discussion will focus on "modern AI," by which I mean statistical machine learning and so-called "deep learning."

1.1. The Magic Hundred Lines of Code

A number of beliefs – some explicit and others implicit – underlie the "modern" approach to AI. First, and perhaps foremost, is the belief that a "magic hundred lines of code" exists that will power a generally intelligent system. Statistical machine learning practitioners have especially fallen prey to this seductive idea. Nearly all AI research is performed by computer scientists, and computer scientists are trained to develop and study the properties of algorithms. The history of the field is littered with "magic hundred lines of code" algorithms, such as AO, Means-End Analysis, Production System interpreters for rule-based expert systems, a host of machine learning algorithms such as ID3, Backpropagation for neural networks, Support Vector Machines, and recent neural network variants such as Recurrent Neural Networks, and Convolutional Neural Networks. Theoretical computer scientists have had a field day studying the properties of these algorithms (and writing papers that 10 people read), but computer science, as a field, has paid too little attention to real-world issues such as understanding real-world problems and user needs, the acquisition and pre-processing of data, and real-world deployment of AI systems. (In fact "true" computer scientists have expressed disdain for focusing on such matters. For example, I was once asked, "What can those of us doing research do for those of you who are doing engineering?" I replied, "Nothing.") Perhaps computer science's sister field of software engineering will expand to address these problems.

1.2. The Unreasonable Effectiveness of Data

Modern AI practitioners generally believe in "the unreasonable effectiveness of data" (see, e.g., [1]). The core of this belief is that running the magic hundred lines of code on the right data will produce an intelligent system that will make good decisions. Even were there a magic hundred lines of code, data of adequate quality and magnitude often does not exist.

Data can be:

- less than required for model building and testing
- low quality (e.g., noisy, incorrectly encoded)
- sparse (e.g., with missing values or missing important training instances; the latter might include instances of the types used in an adversarial machine learning attack)
- inappropriate for the type of problem being solved (e.g., lots of multi-valued nominal variables as input to regression modeling)
- incorrectly/inconsistently labeled (most machine learning algorithms rely on labeled data and assume the labeling is correct and consistent)

Most concerning, however is that data is often biased, either intentionally or unintentionally, often because the data is taken from historical sources. Those biases are then encoded (in black-box form) in the models that machine learning algorithms produce. The resulting AI systems are trusted to make high-stakes decisions, sometimes life-or-death decisions. This is especially risky when their users don't understand how those decisions are made (and the systems in question are not adequately validated and might have been developed using poor methodologies). Examples domains of concern, where AI models are increasingly used, include:

- hiring
- college admissions
- credit approval
- criminal sentencing

(and, coming soon if not already here, lethal attacks by autonomous drones). In many cases, people might trust that the recommendations made by an automated system are neutral and reliable, but AI technologies could simply be a way to "sanitize" biases, rather than to improve biased decision making.

When asked directly if they trust AI (in 2018), 25% of people said "yes", and 28% were neutral [2]. However, approaching the question differently:

- Nearly 60 percent of the 1,000 U.S. managers who responded to the survey think that robots and artificial intelligence perform higher-quality work than humans [3].
- 64% of Workers "trust a robot more than their manager" [4].
- 67% of people would "trust robots more than humans" to manage finance and 59% of people now say they would "trust a robot" to manage finances more than themselves [5].
- significant majorities of people trust Google search results to provide reliable information about financial, legal, and medical decisions [6].

1.3. The Belief that Model Building is What Matters

Based on their university educations, many machine learning engineers believe that model building is the bulk of (and most important part of) the work. It is certainly the sexy part.

Professional data scientists (people working in the real world to solve business and other organizations' problems) know from experience that data extraction, data preprocessing, exploratory data analysis, model evaluation, and deployment make up at least 80% of the work.

From the beginning, it has been the case that obtaining and representing training data has taken the bulk of time in a project. For example, it has been reported anecdotally that Ross Quinlan spent two months representing chess endgames during his seminal ID3 machine learning project; the ID3 algorithm then ran on that data for seven seconds (on a Digital Equipment VAX computer) to produce a model for classifying board positions as won or lost. Modern neural network approaches have had some success with automated feature extraction, but for most applications they too rely on labeled data with all of the attendant problems that human labeling entails, especially when that labeling is being done by untrained people, often working for low wages.

On the back end of the modeling process, machine learning engineers are often not sophisticated about validation, often producing overfit models and sometimes falling prey to the "accuracy paradox" [7].

1.4. The Inability to Understand

A related problem is the inability of an AI system to actually understand its inputs and outputs. For example, a few years ago I input the text from a white power website into such a program in response to being prompted to write a short essay on racial tolerance and received positive feedback. Similarly, MIT researchers created a gibberish essay that easily fooled an automated essay evaluation system [8].

1.5. The Lack of Commonsense

Implementors and their clients implicitly believe that AI systems can function at a high level while lacking the commonsense knowledge that comes from human(-like) experience and thus lacking the ability to reason about novel or atypical input, especially to recognize when that input doesn't make sense, or to generalize reliably from limited input (see, e.g., one-shot learning).

For example, a human can apply commonsense knowledge to recognize a tree appearing in images taken in different seasons as being the same tree, or to recognize an object in a novel orientation. Psychological research has also suggested that humans make use of commonsense knowledge and existing knowledge of previously learned classes to learning new ones efficiently, perhaps from one or only a few instances. (Research on transfer learning in neural networks is attempting to approximate this phenomenon, thus reducing the need for large training datasets. Transfer learning seems to work best when the features encoded in a initial model are general rather than problem specific.)

More importantly, humans have the ability, gained through experience, to recognize when information provided to them is puzzling, ambiguous, or simply doesn't make sense.

1.6. Faith in Benevolent Inputs

Both model building and subsequent system use are typically based on the implicit belief that inputs will be benevolent, not provided by agents who want the system to fail and who will use deceptive input to cause this to happen (see adversarial machine learning). In fact, most application projects probably don't consider this issue at all.

Adversarial attacks on machine-learning-produced models (see, e.g., [9]) are increasingly frequent in domains such as cybersecurity. System developers working in such domains must devote much more attention to anticipating malicious input and training a system to recognize it. Recognizing and addressing such "edge cases" has proven to be difficult for software developers in general; one must believe it will be at least as difficult for machine learning engineers, especially given that knowledgeable adversaries are likely to be actively devising such cases.

1.7. Humans Should Adapt to AI Systems rather than Vice Versa

There are fundamental differences in the ways that humans and AI systems learn and reason. Not knowing how to address these differences, they are often ignored or dismissed as unimportant. For the foreseeable future, deployment will require humans and AI systems to work together. Arguably the systems need to adapt to humans and to respond in human-understandable ways rather than insisting that humans adapt to the systems. We have seen this, by analogy, in the business-process reengineering that was the rage in the 1990's; most such projects failed because they ignored fundamental human factors and expected people to function like machines.

Recently there has been a lot of interest in "explainable AI" in which a system's decision making can be understood by humans; this is likely to prove quite difficult in domains in which (e.g.) neural networks extract very different features than a human would and use those features in a nonhuman decision-making process. 59% of managers don't believe the lack of "explainability" of AI-made decisions to be a major risk [10]. It might prove to be the case that a system's process must be translated into a not precisely accurate, but human-comprehensible, form to generate human-understandable explanations.

1.8. Ethics Matters and AI Systems Lack It

As noted earlier, that AI systems are increasingly being trusted to make decisions with significant ethical dimensions, often by users who don't understand how and why the decisions were reached. As novel example, Jeremy Clarkson, a host of the BBC automotive show Top Gear, posed the following (paraphrased) dilemma for self-driving cars: Suppose a truck swerves into the lane of a self-driving car which is transporting a single elderly person. The only way the car can avoid a head-on collision, which would kill its occupant, is to swerve onto the adjacent sidewalk where a group of children is walking. What should the car do? Consideration of ethics be extended to consideration of who is culpable for damage resulting from poor decisions made by an AI system. Is it the system itself; is it the developer; is it the data provider; and is the liability civil or criminal? Are these issues that technologists, per se, are capable of dealing with as is typically the case at present? If not, who should resolve them and how?

1.9. Even If an AI System Doesn't Do Evil, Will It Do Good?

We must consider if deploying an AI system will actually improve outcomes (and lower costs if that is relevant), or if doing so constitutes unintentional evil. Deploying any type of automation is difficult and costly, and just because you can automate something doesn't mean you should. The recent failure of IBM's Watson as a tool to improve medical care is a major case in point. "IBM has discovered that its powerful technology is no match for the messy reality of today's health care system. And in trying to apply Watson to cancer treatment, one of medicine's biggest challenges, IBM encountered a fundamental mismatch between the way machines learn and the way doctors work." And quoting Martin Kohn, the former chief medical scientist for IBM Research: "Merely proving that you have powerful technology is not sufficient," he says. "Prove to me that it will actually do something useful—that it will make my life better, and my patients' lives better" [11].

2. A Final Question

Are we as practitioners, and more broadly society as a whole, going to address these issues, or are we simply going to say: "We're making a lot of money, so why worry?"

References

[1] A. Halevy, P. Norvig, F. Pereira, The Unreasonable Effectiveness of Data, IEEE Intelligent Systems 24 (2009) 8–12.

[2] Statista, Share of people who agree they trust artificial intelligence in 2018, by country, 2021. URL: https://www.statista.com/statistics/948531/trust-artificial-intelligence-country/.

[3] D. Westermann King, Are Robots Better Employees? Some Managers Think So, Human Resource Executive, 2019. URL: https://bit.ly/3ieV2Pl.

[4] ORACLE, New Study: 64% of People Trust a Robot More Than Their Manager, 2019. URL: https://bit.ly/3xf5qe5.

[5] ORACLE, Global Study: People Trust Robots More Than Themselves with Money, 2021. URL: https://bit.ly/3iiR0pi.

[6] L. Ray, 2020 Google Search Survey: How Much Do Users Trust Their Search Results?, MOZ.com, 2020. URL: https://moz.com/blog/2020-google-search-survey.

[7] T. Afonja, Accuracy Paradox, towardsdatascience.com, 2017. URL: https://bit.ly/3xg2XQB.

[8] L. Perelman, Basic Automatic b.s. Essay Language Generator (BABEL), lesperelman.com, 2014. URL: https://lesperelman.com/writing-assessment-robo-grading/babel-generator/.

[9] OpenAI, Attacking Machine Learning with Adversarial Examples, openai.com, 2017. URL: https://openai.com/blog/adversarial-example-research/.

[10] T. Balakrishnan, M. Chui, B. Hall, N. Henke, Attacking Machine Learning with Adversarial Examples, mckinsey.com, 2020. URL: https://mck.co/3jdI83l.

[11] E. Strickland, How IBM Watson overpromised and underdelivered on AI health care, IEEE Spectrum: Technology, Engineering, and Science News (2019). URL: https://bit.ly/3javL8d.