Logan Jones

20 July 2024

Artificial Intelligence or Information Apocalypse?

An unprecedented rollout of Artificial Intelligence (AI) technology (tech) has taken the world by storm, leaving regulatory bodies, industry, and the public ill-prepared for this phase-transition in how humans interact with tech and information. I aim to show you how some of the current and future states of AI are poised to affect humanity. Due to the breadth of this issue, I will be taking a multi-pronged approach. Offering a background of the current state of AI, its harms, risks, and why the tech is so hard to control, with more detailed highlights on the theme of AI and information literacy. I will add to a literature base that thinks of AI as one of the most pressing concerns humanity faces, and I ask that you take it seriously as well. Without thoughtful prevention these toxic side-effects will become ever more potent as AI matures.

Right away you probably find yourself skeptical, wondering if these concerns could be overblown. The field of AI is notorious for its hype. A very real subset of which has been exaggerated. What sets this apart from other tech predictions? One key component lies in that many of the deleterious effects I will be considering are already here. This is not some far flung future risk. AI, in its current form, is already doing harm. Extrapolating from today, into the near future, is the best we can do given the data that we have. I pull from the findings of multiple fields with a focus on areas of AI harm with broad expert consensus.

The current harms of AI fall on a spectrum from day-to-day inconveniences, needing to cater your resume–and possibly the name you use–to AI filters for fear of auto-rejection (O'neil 113; Bertrand, et al. 1011)., scheduling software that makes last minute alterations leaving

parents scrambling for transportation (O'neil 124-125; Susan, et al. 14)., or the inflationary cost of college tuition (O'neil 50-67). to some of the most important issues humanity is currently facing, the health of the global economy (O'neil 32-48). and even democracy itself (Christiano 1; Schleffer & Miller; Ressa 119-144). The following case illustrates how the creation of AI deep fake content can and will be used to influence elections and spread misinformation. Please keep in mind that the political affiliation considered is completely irrelevant to the issue discussed. The damage can and will cross party–and ideological–lines, it is going to impact all of us.

In Jan. 2024 residents in New Hampshire started receiving phone calls from an AI-generated President Biden warning them not to vote in the primary election or risk bolstering the Trump campaign. Here's a snippet of what the fake Biden said to voters: "Voting this Tuesday only enables the Republicans in their quest to elect Donald Trump again. Your vote makes a difference in November, not this Tuesday" (CTNewsJunkie). I hope you can see that altering what is said or swapping the Biden clip for one of Trump, or any other authority figure for that matter, like a spokesperson for the World Health Organization, is easily within reach. In a Bipartisan response to these robo-calls, here is what Connecticut Attorney General Tong had to say: "This was a deeply disturbing use of artificial intelligence to disrupt and discourage voter participation. If it happened in New Hampshire, we need to assume this will continue to occur again" (Tong). It is easy to imagine misinformation campaigns of this kind happening days or hours prior to a major election or during the response to a serious crisis. Once that genie is out of its bottle it is impossible to fully roll-back the harm. Punishing bad actors after the fact does little to deter the damage to the informed electorate. Preventative measures should be in place to stop AI harm before it occurs. We don't sit around waiting to punish those who would bomb times

square after the bomb has already gone off, why would we wait on AI? This is largely the approach major AI players have been taking, responding after the damage is done.

Oxford AI ethicist Mittelstadt concludes that "Many initiatives, particularly those sponsored by industry, have been characterized as mere virtue-signalling intended to delay regulation and pre-emptively focus debate on abstract problems and technical solutions" (Mittelstadt 1). This description leads us to believe that the current rollout of AI has been controlled by an industry that distracts from the most pernicious areas of harm. Keep this in mind as we take a look at one of the major AI rollout frameworks.

One of the most well-funded–and well intentioned–responses to AI came about in July 2024. We now have the U.S. National Institute of Technologies (NIST) first public draft of "Managing Misuse Risk for Dual-Use Foundation Models" (NIST). While this is a step in the right direction if our goal is a more comprehensive U.S. response to AI. However, I would argue that it is nowhere near sufficient. Firstly, it was designed to inform future policy but is a purely voluntary framework. This means that individual AI developers or companies must "opt-in" to the provided guidelines. That leaves the decision making in the hands of the same industry that Mittelstadt described as not up to the task (1). The second issue lies in its scope, the following discusses the limits of their approach.

> This document does not address other important risks from foundation models, such as bias, discrimination, and hallucination, nor does it address all risks to public safety, including those that may arise from other types of AI models and systems (NIST 3).

With this limited U.S. approach in mind, let us consider some of the other players.

Research shows that all the major national players are taking a similar approach to the rollout of AI. "key findings suggest a surprising consistency in the narrative of these strategies,

converting bold and vague policy talk into a seemingly inevitable technological pathway" (Bareis and Katzenbach 856). These findings go to show that the aforementioned national players see the rollout of disruptive AI technologies as a foregone conclusion. It is no longer a question of if, it is a question of how much and how soon. With each player vying to maintain the tech advantage, preventing future loss or harm is an afterthought at best. Given this sense of inevitability, two pathways materialize. The one we are on, where the technological wave crashes over us and another solution where we cautiously direct its flow to minimize harm. The following scenario shows how difficult it is to be sure you are on the safer technological path.

In a human-centered approach to showcasing "The Alignment Problem" an anecdote from Brian Christian portrays how difficult it is to translate what we humans want, our values, into appropriate technological solutions, even with something as simple as heating our homes. Christian details waking up on a cold New England night, drenched in sweat, fearing his house is on fire. He rushes out the door to find a house that is peaceful, dark and cold. At this point realization starts to dawn on him, the spare bedroom with the thermostat in it was open to the frigid temperatures while his bedroom had been closed off. No matter how much heat the system pumped into the two bedrooms, the thermostat never registered an increase in temperature. This resulted in the bedroom he and his wife were in to be heated to an uncomfortable degree (Christian 311). Considering this story with respect to AI Joseph Taylor remarked "AI is like having a million thermostats with no idea what each of them does." This conveys the magnitude of the Alignment Problem with respect to AI in a similar way to Christian himself, "A growing chorus within the AI community–first a few voices on the fringe, and increasingly the mainstream of the field–believes, if we are not sufficiently careful, that this is literally how the world will end" (Christian 10).

One of the most concerning aspects of AI tech and its relation to information literacy comes from so-called Black Box Models (black box), an easier way to think about black boxes lies in an analogy to the thermostat story. If the air conditioning system was an AI software and the bedroom with the thermostat was data inaccessible to developers–in a way that they could not see why it was continually spitting out hot air–it would be considered a black box. They solve computational problems but us humans on the outside have no way to know how they solved the problem or why that particular solution was chosen. For some kinds of AI this does not carry much risk, who cares how it came up with the spaghetti recipe as long as it tastes good right? For other models, like those used in criminal justice or healthcare decision making, the results can be far more sinister.

I have seen troubling examples of black boxes that are otherwise as accurate as medical professionals at flagging malignant tumors, flagging any photos with measuring rulers present (Narla, et al). As well as criminal justice recidivism software that, “Has been given to thousands of inmates since its invention in 1995” and uses approximate information or proxies to sentence people differently based on “who they are” rather than purely on the crime they have committed as our justice system intends (O’neil 23-27). Apart from a lack of information into how these AI’s work, let's take a deeper look at how they are hindering public education.

I think most of us agree that information literacy is something to strive towards. Being able to discern between what is real and what is fake is fundamental to being a successful human. From what is safe to eat or drink without getting sick to who to vote into democratic office, having accurate information is the necessary first step towards making good decisions. One critical aspect of information literacy is education. Information literacy is not something that all people are born highly proficient in. It is a skill that must be learned and developed.

In a systematic review “The New Reality of Education in the Face of Advances in Generative Artificial Intelligence” the authors take a look at how the current generation of large language models like Chat GPT are already re-shaping academia (García-Peñalvo, et al). Luckily, they found a general consensus among experts that the use of these tools should not be prohibited, because they also found that there is no known relevant case in the literature of a tech of this kind being successfully prohibited. What this means in effect, is that it may not be possible to prevent the use of these technologies. They concluded their review with the following “we have realized our main challenge: the speed at which we will have to analyze and incorporate these innovations” (García-Peñalvo, et al).

This is a direct example of a readily available tech in the form of the current edition of transformers being impossible to contain in practice. This leads us to consider whether we can at least track their use and differentiate artificial content from human content? “Testing of Detection Tools for AI-Generated Text.” With its focus on generated text again suggests that no, we can’t reliably differentiate between human and artificial content today, they go on to show that as the tech matures our ability to differentiate the two is likely to get worse going forward (Weber-Wulff, et al).

If we cannot know what is made by a human and what is made by a “bot”, how are we supposed to approach and evaluate that information? I’ll be the first to admit that–as with most of the previously discussed issues–there are no easy answers. A crucial first step is to collectively take this problem seriously. We need to quickly learn how to apply the methods we use today to evaluate human generated content, to an exponentially growing subset of information, artificially generated information. As these technologies mature it won't just be text content that we need to worry about. Photorealistic video with voice generated deep-fake sound design–like the

robo-calls we saw earlier–will be combined in the next generation of AI tools. This will irreversibly alter the information landscape as we know it. In an age of fake news and divisive politics it's certainly cause for concern.

There is still some good news, AI is not inherently bad. AI is already doing a lot to enrich the lives of humans. The tech is in its infancy and if we choose the right path now, we can minimize a significant amount of the coming harm while still benefiting from AI. One way that individuals can help is by doing exactly as you are, informing yourself of the risks of AI through the lens of expert consensus. By staying informed and discussing these topics, we position society to better effect the changes we need. A powerful tool for this is through political action. Make sure you are voting for representatives who take AI and expert opinion seriously and hold them accountable for their AI policy. Do your best to boycott products and companies that you feel are handling AI problems like containment, black boxes or content generation in dangerous ways. I hope that at the very least you have inherited a newfound sense of urgency for combating the risks of AI. If we can leverage this powerful new technology in a way that reduces its harmful side effects, AI will have a lasting legacy as one of the most beneficial technologies that humans have ever created. I do not ask for you to throw out this vision wholesale, I humbly ask that you consider taking a better informed, slower, and safer approach to the rollout of Artificial Intelligence.

Works Cited

Bareis, Jashcha and Katzenbach, Christian. “Talking AI into Being: The Narratives and imaginaries of National AI Strategies and Their Performative Politics” *Science, Technology, & Human Values,* vol. 47, no. 5, 2022, pp. 855-881, https://doi.org/10.1177/01622439211030007

Bertrand, Susan, et al. “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review,* vol. 94, no. 4, Aug. 2004, pp. 991–1013, https://doi.org/10.1257/0002828042002561.

Christian, Brian. The Alignment Problem: Machine Learning and Human Values. W.W. Norton & Company, 2021.

Christiano, Thomas. "Algorithms, Manipulation, and Democracy." *Canadian Journal of Philosophy,* vol. 52, no. 1, 2022, pp. 109-124. doi:https://doi.org/10.1017/can.2021.29.

CTNewsJunkie. “Biden - Deepfake.” SoundCloud, SoundCloud, 2024, soundcloud.com/ctnewsjunkie/biden-deepfake?utm_source=ctnewsjunkie.com&utm_campaign=wtshare&utm_medium=widget&utm_content=https%253A%252F%252Fsoundcloud.com%252Fctnewsjunkie%252Fbiden-deepfake.

García-Peñalvo, Francisco José, et al. "The New Reality of Education in the Face of Advances in Generative Artificial Intelligence" *Revista Iberoamericana De Educación a Distancia,* vol. 27, no. 1, 2024, pp. 9-32. https://doi.org/10.5944/ried.27.1.37716.

Lambert, Susan, et al. "Precarious Work Schedules among Early-Career Employees in the US:

National Snapshot." *Research brief,* University of Chicago, Chicago, IL 2014.

Mittelstadt, Brent. “Principles Alone Cannot Guarantee Ethical AI.” *Nature Machine Intelligence,* vol. 1, no. 11, Nov. 2019, pp. 501–07, https://doi.org/10.1038/s42256-019-0114-4.

Narla, Akhila, et al. “Automated Classification of Skin Lesions: From Pixels to Practice.” *Journal of Investigative Dermatology,* vol. 138, no. 10, Oct. 2018, pp. 2108–10, https://doi.org/10.1016/j.jid.2018.06.175.

O’neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books, 2016.

Ressa, Maria. *How to Stand up to a Dictator: The Fight For Our Future*. HarperCollins, 2022.

Schleffer, Guy, and Benjamin Miller. “The Political Effects of Social Media Platforms on Different Regime Types.” *Texas National Security Review*, 2021, tnsr.org/2021/07/the-political-effects-of-social-media-platforms-on-different-regime-types/.

Taylor, Joseph. Peer Review Feedback. Email, 6 Aug. 2024.

Team, NIST AIRC. “NIST AIRC - Home.” NIST Trustworthy & Responsible AI Resource Center, 26 Jan. 2023, https://airc.nist.gov/Home. Accessed 6 July 2024.

Tong, William. “Attorney General Tong Issues Bipartisan Warning to Suspected Election Scam AI Robocallers.” CT.gov, 2024, portal.ct.gov/ag/press-releases/2024-press-releases/attorney-general-tong-issues-bipartisan-warning-to-suspected-election-scam-ai-robocallers.

Weber-Wulff, Debora, et al. “Testing of Detection Tools for AI-Generated Text.”

International Journal for Educational Integrity, vol. 19, no. 1, 2023, pp. 1-39. https://doi.org10.1007/s40979-023-00146-z.