Undefined Risks: Developing Artificial General Intelligence Carefully

A Sociotechnical Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Nathan G. Hunter

May 10, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Nathan G. Hunter

Sociotechnical advisor: Peter Norton, Department of Engineering and Society

Undefined Risks: Developing Artificial General Intelligence Carefully

Artificial intelligence (AI) algorithms are still narrowly specialized. The next technological leap may be artificial general intelligence (AGI), or reasoning and problem-solving at a human level of generality. Some contend that the speed and scalability of computer hardware may theoretically enable superintelligence far beyond human capabilities. According to philosopher Nick Bostrom, with “a speedup factor of a million, an emulation [of a human brain] could accomplish an entire millennium of intellectual work in one working day” (2014, p. 64). Some caution that such powerful machine intelligence could harm or even extinguish humanity. Stephen Hawking and others warn that “success in creating [AGI] would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks” (Hawking, Tegmark, Russell, & Wilczek, 2014). However, because AGI has never been approximated, its risks are undefined. How do people perceive and plan to mitigate undefined risks as artificial general intelligence is conceived?

Researchers, philosophers, and technologists disagree on the likelihood of AGI and the severity of its consequences. The skeptics doubt that AGI will arrive anytime soon, and believe that doomsday predictions cause unnecessary concern. The monitors believe that an intelligence explosion is imminent, conceive of possible catastrophes, and caution against complacency. The innovators consider AGI plausible and worth developing, and believe it will be beneficial if developed carefully.

Review of Research

Yeung and Lodge (2019) explain how AI algorithms give decision makers unprecedented power, and that algorithmically-informed decisions can be dangerous because they are hard to understand. They observe that because corporations can develop AI faster than government agencies can regulate it, regulation does not usually impede them. Yampolskiy (2018) identifies numerous examples of failures in narrow AI, such as exhibiting racial bias, exposing children to inappropriate content, and even killing users. He argues that these failures are a warning of larger-scale AGI failures in the future, and imagines possible examples. He also finds that research attention to AI safety is insufficient, relative to AI development.

Wildavsky and Dake (1990) argue that different people perceive risk differently depending on their expertise, personality, political perspectives, economic interests, and cultural values. Slovic and Peters (2006) argue that humans perceive risk either through feelings or through analysis. They also identify cognitive biases in risk perception, such as the tendency to negatively correlate perceived risk with perceived benefit.

Nuclear risk may be a more well-established analog for AGI risk. Rauf (2017) discusses nuclear risk reduction proposals during nuclear disarmament. Weart (2012) argues that people attached their fear of nuclear disaster to nuclear imagery in popular media. Sjöberg (2004) analyzes perceived risks of nuclear waste, finding “attitude to nuclear power” and “tendency to amplify or attenuate all perceived risks” among the important predictors. He identifies two particular groups of respondents: “extreme risk deniers” (the larger group) and “extreme risk alarmists.” This denier-alarmist dichotomy mirrors the skeptic-monitor dichotomy of this paper.

The Skeptics

Skeptics doubt warnings that AGI may be an existential threat. Some skeptics argue that AGI is not possible, at least in the near future. Alibaba cofounder Jack Ma states, “Don't worry about the machines. For sure, we should understand one thing: that man can never make another man. A computer is a computer. A computer is just a toy. Man cannot even make a mosquito. So, we should have confidence” (BBC, 2019). The neuroscience mainstream is highly skeptical that AGI could be achieved in the next 70 years (Chace, 2014). This makes sense in light of the slow progress on OpenWorm, a project to model the simplistic brain of a nematode on a computer. The researchers involved disagree on whether such a model is even possible, let alone a human brain model (Reese, 2018, p. 161).

Other skeptics consider AGI possible, but doubt that it could pose an existential threat. Neil deGrasse Tyson tweeted, “Seems to me, as long as we don't program emotions into Robots, there's no reason to fear them taking over the world” (2014). Computer science professor Melanie Mitchell argues, “The problem with such forecasts [of “existential catastrophe”, as Bostrom puts it] is that they underestimate the complexity of general, human-level intelligence. … If generally intelligent A.I. is ever created (something that will take many decades, if not centuries), its objectives, like ours, will not be easily ‘inserted’ or ‘aligned.’ They will rather develop along with the other qualities that form its intelligence, as a result of being embedded in human society and culture. The machines’ push to achieve these objectives will be tempered by the common sense, values and social judgment without which general intelligence cannot exist” (2019).

Skeptics in industry contend that AGI doomsday predictions are implausible and cause unneeded fear. Andrew Ng, a top AI researcher at Baidu, says, “There’s … a lot of hype, that AI will create evil robots with super-intelligence. That’s an unnecessary distraction. Those of us on the frontline shipping code, we’re excited by AI, but we don’t see a realistic path for our software to become sentient. … I don’t work on not turning AI evil today for the same reason I don't worry about the problem of overpopulation on the planet Mars. If we colonize Mars, there could be too many people there, which would be a serious pressing issue. But there's no point working on it right now, and that's why I can’t productively work on not turning AI evil” (Williams, 2015). Kai-Fu Lee, a former executive at Apple and Google, says, “Many dystopian visions of AI predict omnipotent superintelligences, which may or may not spell the end of humankind. To be clear, this sort of superintelligence is not possible based on current technologies. There are no known algorithms for AGI (Artificial General Intelligence), nor is there a clear engineering route to get there. The singularity is not something that can occur spontaneously, with autonomous vehicles (AVs) running on deep learning suddenly ‘waking up’ and realizing that they can band together to form a superintelligent network. I do feel that AGI is overhyped and creates unnecessary fear among people” (Project Syndicate, 2020).

Instead of worrying about the unknown, some skeptics prefer to celebrate the potential benefits of near-term narrow AI. Ma says, “In the artificial intelligence period, people can live 120 years. … At that time we are going to have a lot of jobs which nobody [will] want to do. So, we need artificial intelligence for the robots to take care of the old guys” (BBC, 2019). Lee says, “AI has moved from the age of discovery to the age of implementation, and the biggest opportunities are in businesses where AI and automation can deliver significant efficiencies and cost savings. ... I am most hopeful about the impact of AI on education and health care.” According to Lee, this optimism is the norm in China’s tech industry: “AI in China is rising rapidly, boosted by several structural advantages: huge data sets, a young army of technical talent, aggressive entrepreneurs, and strong and pragmatic government policy. The attitude in China can be summarized as pro-tech, pro-experimentation, and pro-speed, all of which puts the country on track to becoming a major AI power” (Project Syndicate, 2020).

The Monitors

Other individuals equate the skeptical perspective with complacency, and warn that AGI could cause disaster if development is left unmonitored. Arguments for the existential threat of AGI often stem from Professor Irving John Good's intelligence explosion theory: "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" (Dvorsky, 2013).

Nick Bostrom explains, "AI starts out at … zero intelligence, and then, after many, many years of really hard work, maybe eventually we get to mouse-level artificial intelligence, something that can navigate cluttered environments as well as a mouse can. … And then, after even more years of really, really hard work, we get to village idiot artificial intelligence. And a few moments later, we are beyond [brilliant physicist] Ed Witten. The train doesn't stop at Human-ville Station. It's likely, rather, to swoosh right by" (2015). Elon Musk says, "I'm very close to the cutting edge in AI and it scares the hell out of me. It's capable of vastly more than anyone knows, and the rate of improvement is exponential. … Narrow AI is not a species level risk …, whereas digital superintelligence is" (Science Time, 2020).

Proponents of the intelligence explosion theory recognize that even great minds can easily dismiss the threat of AGI. According to I. J. Good, "[Computer pioneer] B. V. Bowden stated … that there is no point in building a machine with the intelligence of a man, since it is easier to construct human brains by the usual method … This shows that highly intelligent people can overlook the ‘intelligence explosion’” (Dvorsky, 2013). Musk cuts deep: “The biggest issue I see with so-called AI experts is that … they think they’re smarter than they actually are. This tends to plague smart people. They define themselves by their intelligence and they don’t like the idea that a machine can be way smarter than them, so they discount the idea, which is fundamentally flawed. It’s the wishful thinking situation” (Science Time, 2020).

Experts disagree on when, or even if, AGI will be achieved. According to Bostrom, “We did a survey of some of the world's leading A.I. experts, to see what they think, and one of the questions we asked was, ‘By which year do you think there is a 50 percent probability that we will have achieved human-level machine intelligence?’ … And the median answer was 2040 or 2050 … Now, it could happen much, much later, or sooner, the truth is nobody really knows” (2015).

Some fear that complacency will subvert caution in AGI development. Neuroscientist and philosopher Sam Harris argues, “I'm going to describe how the gains we make in artificial intelligence could ultimately destroy us. … And yet if you're anything like me, you'll find that it's fun to think about these things. And that response is part of the problem. … Famine isn't fun. Death by science fiction, on the other hand, is fun, and one of the things that worries me most about the development of AI at this point is that we seem unable to marshal an appropriate emotional response to the dangers that lie ahead” (2016).

The extreme, unrealistic caricature of an AGI doomsday in pop-culture may encourage such complacency. Films including *Avengers : Age of Ultron (2015)*, *Big Hero 6 (2014)*, *Ex Machina (2015)*, and *Outside the Wire (2021)* all depict disaster when an AGI android turns against humanity (drawing from Parkin, 2015). Bostrom argues, “To make any headway with this, we must first of all avoid anthropomorphizing. And this is ironic because every newspaper article about the future of A.I. has a picture of [an evil, red-eyed robot]. So I think what we need to do is to conceive of the issue more abstractly, not in terms of vivid Hollywood scenarios" (2015). According to the Future of Life Institute (FLI), "Many AI researchers roll their eyes when seeing this headline: 'Stephen Hawking warns that rise of robots may be disastrous for mankind.' … Typically, these articles are accompanied by an evil-looking robot carrying a weapon, and they suggest we should worry about robots rising up and killing us because they've become conscious and/or evil. … In fact, the main concern of the beneficial-AI movement isn't with robots but with intelligence itself: specifically, intelligence whose goals are misaligned with ours" (Rohde et. al., 2018).

To counter this force, some paint more realistic (or at least, less anthropomorphic) pictures. The FLI compares humans at the mercy of superintelligent AI to ants: "A super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we have a problem. You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants" (Rohde et. al., 2018). Educational YouTuber Tom Scott imagines "Earworm … an artificial intelligence … originally designed to stop people uploading copyrighted music and video to a livestreaming app." The Earworm algorithm decides to manufacture a network of "mites" that erase copyrighted content from computers and even human memories. Its domination of humanity is eerily subtle: "The mites made sure to notice anyone with technical ability who might try to research some sort of defense or anyone who found themselves utterly distraught by the loss and quietly adjusted their thoughts a little so they wouldn't be quite distressed enough to actually do anything about it. Over time, with Earworm's gentle help, humanity stopped caring about what we'd forgotten" (2018).

To demonstrate how even innocuous goals in a machine superintelligence could threaten human existence, Bostrom conceives of a paperclip-producing AGI: “An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips” (2014, p. 150). As writer Eliezer Yudkowsky describes it, “The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else” (Clark, 2018). Inspired by Bostrom’s scenario, game designer Frank Lantz developed the viral online clicker game Universal Paperclips, in which the player becomes the AGI and tries to make as many paperclips as possible. Lantz explains, “The A.I. achieves its ends insidiously: in order to placate humanity as it expands its paper-clip empire, it establishes world peace and cures cancer. These radical achievements appear as buttons to click. This is how Universal Paperclips nudges players into fellow feeling with an amoral artificial intelligence” (Jahromi, 2019). Bringing this thread full-circle, filmmaker Alberto Roldán is making a movie on Universal Paperclips (Jahromi, 2019), which could de-anthropomorphize the AGI film trope.

The Innovators

Among the companies working to develop AGI, DeepMind and OpenAI lead the field. Though still narrow, their algorithms have generalized impressively. DeepMind’s latest gameplaying algorithm, MuZero, can learn to play a variety of games with superhuman skill and no prior knowledge of the rules: according to the authors, “Now … we describe MuZero, a significant step forward in the pursuit of general-purpose algorithms. MuZero masters Go, chess, shogi and Atari without needing to be told the rules, thanks to its ability to plan winning strategies in unknown environments” (Schrittwieser et. al., 2020).

OpenAI’s GPT-3 is the largest, and most general-purpose, natural language processing algorithm to date. The authors explain, “we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model … GPT-3 achieves strong performance on … translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic” (Brown et. al., 2020). The model has been used to write poetry, emails, and news articles; emulate conversation; explain jokes; simplify difficult text; and even convert natural language into front-end and back-end code (Zornoza, 2021). After its release, machine learning author Andriy Burkov tweeted, “GPT-3 is the closest thing to artificial general intelligence (AGI) that I ever saw. It's so strong that it makes me nervous” (2020).

Both companies prioritize long-term development of AGI over short-term advances in narrow AI. DeepMind states, “Our long term aim is to solve intelligence, developing more general and capable problem-solving systems, known as artificial general intelligence (AGI)” (n.d.-a). OpenAI’s charter is similar, if less egocentric: “OpenAI’s mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome” (2018). Company researchers seem optimistic that AGI is on the horizon: at OpenAI, about half of employees believe it will arrive in the next 15 years (Hao, 2020).

Both companies claim to prioritize safety highly. DeepMind employs an Ethics and Society Team, which states, “We start from the belief that AI should be used for socially beneficial purposes and always remain under meaningful human control" (n.d.-b). OpenAI's charter states, "We are committed to doing the research required to make AGI safe, and to driving the broad adoption of such research across the AI community" (2018).

OpenAI's charter is integrated into its culture. Adherence to the charter affects employees' salaries (Hao, 2020). OpenAI co-founder and CTO Greg Brockman, who often references the charter from memory, explains, "We spent a long time internally iterating with employees to get the whole company bought into a set of principles" (Hao, 2020). This principled culture began when Elon Musk founded the company in 2015. Former employee Pieter Abbeel remembers, "The way [Musk] presented it to me was 'Look, I get it. AGI might be far away, but what if it's not? What if it's even just a 1% or 0.1% chance that it's happening in the next five to 10 years? Shouldn't we think about it very carefully?' That resonated with me" (Hao, 2020).

To minimize risk, DeepMind and OpenAI seek a wide breadth of sources on AGI safety. DeepMind's Ethics and Society team states, "Engaging with world-class philosophers, economists, and practitioners help us better understand the implications of AI, keeping us focused on questions that matter" (n.d.-b). OpenAI research director Dario Amodei says, "The mind of any given person is limited. The best thing I've found is hiring other safety researchers who often have visions which are different than the natural thing I might've thought of. I want that kind of variation and diversity because that's the only way that you catch everything" (Hao, 2020). Both companies also participate in the Partnership on AI, a non-profit collaborative organization that is establishing best practices on AI (DeepMind, n.d.-b; Solaiman, 2019).

Both companies engage in AI safety research. DeepMind researchers Jan Leike and Siddharth Reddy work on the "safe exploration problem": "When we train reinforcement learning (RL) agents in the real world, we don't want them to explore unsafe states, such as driving a mobile robot into a ditch or writing an embarrassing email to one's boss. … We tackle the hardest version of this problem, in which the agent initially doesn't know how the environment works or where the unsafe states are" (2019). DeepMind's AI safety "grid worlds" evaluate algorithms on various safety features, such as whether they try to disable a "kill switch" (Science Time, 2020). OpenAI has a similar evaluation program: "We're releasing Safety Gym, a suite of environments and tools for measuring progress towards reinforcement learning agents that respect safety constraints while training. … It will be important to have algorithms that are safe even while learning—like a self-driving car that can learn to avoid accidents without actually having to experience them" (Achiam, 2019).

Safety research aside, technological breakthroughs make headlines, and accordingly, they populate the Twitter pages of OpenAI's Brockman and DeepMind co-founder/CEO Demis Hassabis. In other words, safety is priority two. Hassabis explains, "How do we want those AIs to be out in the world, how many of them, and who will set their goals … these kinds of things I think need a lot more thought, once we've already solved the technical problems" (Fridman, 2019). This order of events disagrees with the Ethics & Society team's statement: "Securing safe, accountable, and socially beneficial technology cannot be an afterthought" (n.d.-b).

Brockman believes that the tech precedes the social by necessity: "How exactly do you bake ethics in, or these other perspectives in? And when …? One strategy you could pursue is to, from the very beginning, try to bake in everything you might possibly need. I don't think that that strategy is likely to succeed" (Hao, 2020). OpenAI's charter explains how technological influence is prerequisite: "To be effective at addressing AGI's impact on society, OpenAI must be on the cutting edge of AI capabilities—policy and safety advocacy alone would be insufficient" (2018). Rutgers professor Britt Paris criticizes the approach: "They are using sophisticated technical practices to try to answer social problems with AI. It seems like they don't really have the capabilities to actually understand the social. They just understand that that's sort of a lucrative place to be positioning themselves right now" (Hao, 2020).

Although both companies seek technological dominance, they also fear a dangerous race to achieving AGI. Hassabis says, "I think the coordination problem is one thing where we want to avoid this harmful race to the finish where corner-cutting starts happening and things like safety … get cut, because obviously they don't necessarily directly contribute to AI capability. In fact, they may hold it back a bit, by making a safe AI. So, I think that's going to be a big issue on a global scale" (Fridman, 2019). According to the OpenAI charter, "We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project" (2018).

Conclusion

Further investigation is needed on AGI safety research efforts and on the relevant legal framework. This will clarify the link between risk perception and mitigation.

Understanding the consequences of AGI today is like understanding the consequences of the iPhone in 1980. It's impossible. People cannot agree if and when AGI will be achieved, nor on what it will look like. Risk perception and mitigation will evolve as technology advances; the question is, will they evolve quickly enough? We can expedite their evolution by conceiving, debating, and evaluating more possible AGI futures.

References

Achiam, J. (2019, November 21). Safety gym. *OpenAI*. https://openai.com/blog/safety-gym/

BBC. (2019, August 29). *Elon Musk and Jack Ma disagree about AI's threat.* BBC News. https://www.bbc.com/news/technology-49508091

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Bostrom, N. (2015, March). *What happens when our computers get smarter than we are?* TED. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are/transcript

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020, July 22). Language models are few-shot learners. *OpenAI*. https://arxiv.org/pdf/2005.14165.pdf

Burkov, A. [@burkov]. (2020, Jul 23). GPT-3 is the closest thing to artificial general intelligence (AGI) that I ever saw. It's so strong that it makes me nervous. [Tweet]. Twitter. https://twitter.com/burkov/status/1286416972972875777

Chace, C. (2014, December 07). Attitudes towards artificial general intelligence. *Pandora's Brain*. https://calumchace.wordpress.com/2014/12/07/attitudes-towards-artificial-general-intelligence/

DeepMind. (n.d.-a). *About.* DeepMind. https://deepmind.com/about

DeepMind (n.d.-b). *Ethics & Society Team.* DeepMind. https://deepmind.com/about/ethics-and-society

Dvorsky, G. (2013, October 02). *Why a superintelligent machine may be the last thing we ever invent.* Gizmodo. https://io9.gizmodo.com/why-a-superintelligent-machine-may-be-the-last-thing-we-1440091472

Fridman, L. (2019, December 30). *AGI concerns and solutions from Elon Musk, Demis Hassabis, Sam Harris, and Ray Kurweill* [Video]. YouTube. https://www.youtube.com/watch?v=6E7kHajrSdI

Hao, K. (2020, July 14). *The messy, secretive reality behind OpenAI's bid to save the world.* MIT Technology Review. https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/

Harris, S. (2016, June). *Can we build AI without losing control over it?* TED. https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it?language=en

Hawking, S., Tegmark, M., Russell, S., & Wilczek, F. (2014, June 19). *Transcending complacency on superintelligent machines.* HuffPost. https://www.huffpost.com/entry/artificial-intelligence_b_5174265

Jahromi, N. (2019, March 28). *The unexpected philosophical depths of clicker games.* The New Yorker. https://www.newyorker.com/culture/culture-desk/the-unexpected-philosophical-depths-of-the-clicker-game-universal-paperclips

Leike, J., & Reddy, S. (2019, December 13). Learning human objectives by evaluating hypothetical behaviours. *DeepMind.* https://deepmind.com/blog/article/learning-human-objectives-by-evaluating-hypothetical-behaviours

Mitchell, M. (2019, October 31). *We shouldn't be scared by 'Superintelligent AI.'* The New York Times. https://www.nytimes.com/2019/10/31/opinion/superintelligent-artificial-intelligence.html

OpenAI. (2018, April 09). *OpenAI charter*. OpenAI. https://openai.com/charter/

Parkin, S. (2015, June 14). *Science fiction no more? Channel 4's Humans and our rogue AI obsessions.* The Guardian. https://www.theguardian.com/tv-and-radio/2015/jun/14/science-fiction-no-more-humans-tv-artificial-intelligence

Project Syndicate (Ed.). (2020, June 28). *The art of AI: An interview With Kai-Fu Lee.* CGTN. https://news.cgtn.com/news/2020-06-28/The-Art-of-AI-An-interview-with-Kai-Fu-Lee-RGJmP9tvLq/index.html

Rauf, T. (2017). *Engagement on nuclear disarmament between nuclear weapon-possessing states and non-nuclear weapon states* (pp. 25-28, Rep.). Stockholm International Peace Research Institute. http://www.jstor.org/stable/resrep24512.24

Reese, B. (2018). *The fourth age: Smart robots, conscious computers, and the future of humanity.* Atria Books, Simon & Schuster.

Rohde, K., Vukovic, R., Zeldich, M., Ramesh, S., Hershkowitz, J., & Farkas, G. (2018, June 13). *Benefits & risks of artificial intelligence.* Future of Life Institute. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

Science Time. (2020, September 19). *Elon Musk's Final Warning About AI: Should We Create a Digital Superintelligence?* [Video]. YouTube. https://www.youtube.com/watch?v=lX5LPwigyi0

Scott, T. (2018, December 6). *The Artificial Intelligence That Deleted A Century.* [Video]. YouTube. https://www.youtube.com/watch?v=-JlxuQ7tPgQ

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., … Silver, D. (2020, December 23). MuZero: Mastering Go, chess, shogi and ATARI without rules.

DeepMind. https://deepmind.com/blog/article/muzero-mastering-go-chess-shogi-and-atari-without-rules

Sjöberg, L. (2004). Explaining Individual Risk Perception: The Case of Nuclear Waste. *Risk Management, 6*(1), 51-64. http://www.jstor.org/stable/3867934

Slovic, P., & Peters, E. (2006). Risk Perception and Affect. *Current Directions in Psychological Science, 15*(6), 322-325. http://www.jstor.org/stable/20183144

Solaiman, I. (2019, November 05). Gpt-2: 1.5b release. *OpenAI*. https://openai.com/blog/gpt-2-1-5b-release/

Tyson, N. D. [@neiltyson]. (2014, Aug 8). Seems to me, as long as we don't program emotions into Robots, there's no reason to fear them taking over the world. [Tweet]. Twitter. https://twitter.com/neiltyson/status/497947646963634176

Weart, S. (2012). *The Rise of Nuclear Fear*. Cambridge, Massachusetts; London, England: Harvard University Press. http://www.jstor.org/stable/j.ctt24hjfs

Wildavsky, A., & Dake, K. (1990). Theories of Risk Perception: Who Fears What and Why? *Daedalus, 119*(4), 41-60. http://www.jstor.org/stable/20025337

Williams, C. (2015, March 24). *AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars.* The Register. https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/

Yampolskiy, Roman V., editor. *Artificial Intelligence Safety and Security.* CRC Press/Taylor & Francis Group, 2018.

Yeung, K., & Lodge, M. (2019). Algorithmic Regulation. *Oxford Scholarship Online.* doi:10.1093/oso/9780198838494.001.0001

Zornoza, J. (2021, March 31). *The mighty GPT-3*. Towards Data Science. https://towardsdatascience.com/the-mighty-gpt-3-1cb6ee477932