Using Prediction Markets to Assert the Danger of AI

Section I: Introduction

Though becoming ever more prevalent in discussions of the future of AI, the academic field of ethics surrounding an "AI pause" remains undeveloped. Some prominent experts on AI including Eliezer Yudkowsky have warned that the risk AI could pose to all of humanity in the future should warrant a shutdown of all research on superintelligent AI. Others in the community have argued that aligning a superintelligent AI with human interests is not a large issue and that AI development should continue at its current pace. In this paper, I will argue that there is an existential threat posed by AI and it is necessary to proceed with caution in AI research by instituting a pause on further AI development.

In section II of this paper, I will discuss some contextual information necessary for my argument. This will include some context about the current positions on the danger of AI, and some information about the predictive power of betting markets. Upon providing some context, I will advance my argument that AI does pose an existential threat to humanity as evidenced by predictive betting markets in Section III. I will conclude in Section IV summarizing my argument in favor of AI being perceived as a danger to humanity and will outline some next steps that can be taken in response to this danger.

Section II: Contextual Information

Before arguing my position on the dangers of AI and ethics of AI, it is important to contextualize perceived dangers of AI and define some key concepts my analysis will rely on. I will first provide the former. One way in which AI may pose a danger is the view subscribed to by Eliezer Yudkowsky that AI research will face an "intelligence explosion" in which AI

becomes so intelligent it is essentially no longer in the control of humans. Kelsey Piper motivates the ways in which such an "intelligence explosion" is dangerous in "A Field Guide to AI Safety. In summary, an AI has the ability to be used to create more AI, and it is currently the case that researchers use AIs to help produce new AIs at a faster rate. As the intelligence of these AIs surpasses that of humans, they may create further AIs that surpass human intelligence by even larger metrics than the previous AIs.[1] Without enough preparation and planning in how to control these superintelligent AIs, it will be hard for humanity to control AI and make sure that AIs are properly aligned with humans. This could result in AI viewing humans not as living things deserving of agency and respect, but as mere atoms to be used however it pleases.[2] An example of what this could result in is an AI creating super bacteria to infect humans and use them for a more "efficient" end.[3] This would quite likely lead to the end of humanity.

Another view which is a little less pessimistic–though notably still viewing AI as possibly dangerous–is that associated with Open Philanthropy. This view still entails the idea of AI creating other AI which far surpasses human intelligence. Instead of positing that this results in the end of humanity, the Open Philanthropy view is that the reinforced learning AI receives in the form of human feedback on answers will only lead to AI that says what it knows humans will approve of and not display what AI would actually choose to do.[4] This lack of alignment faces the risk of growing as AI models become more and more intelligent and less subject to human surveillance on their methods. If humans then rely on AI in areas–such as labor for example–it is

---

[1] Kelsey Piper, "A Field Guide to AI Safety," *Asterisk Magazine*, June, 2023, https://asteriskmag.com/issues/03/a-field-guide-to-ai-safety#the-intelligence-explosion-and-early-work-on-preventing-ai-catastrophe.
[2] Yudkowsky, Eliezer. "Pausing AI Developments Isn't Enough. We Need to Shut it All Down." *Time Magazine*, March 29, 2023. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.
[3] Piper, "AI Safety."
[4] *Ibid*.

very possible that AI will do what it wants and not what humans want.[5] Thus, humans would be handing over control of the world to AI without even realizing it.

Some individuals, including those working for companies such as Meta, Google, and OpenAI, seem to hold the belief that AI does not pose an existential risk to humanity. This is rooted in their view that the alignment problems identified by other experts should not be as difficult to solve as many think. Rather, they believe that the work needed to make AI commercially useful will also result in AI being aligned enough for it to pose no threat to humanity.[6]

Another development has been the stance of several AI experts against further development without a mitigation of risk. For example, the Center for AI Safety created a statement calling for the need to treat AI risk like other existential risks like pandemics and nuclear wars.[7] Several notable AI experts appear on the list of signatories such as Yoshua Benjio, Sam Altman, and Dario Amodei.[8] This illustrates a general belief that AI risk should be mitigated by those working on AI.

One key concept it is important to understand for my analysis is the use of prediction markets. A prediction market, as described by Scott Alexander, is like a stock market "for beliefs about future events" where the price at which people buy shares equates to the belief that the event will or will not happen.[9] Alexander holds that prediction markets are generally as accurate as any expert in a given field, since otherwise this would result in someone taking

---

[5] *Ibid*.
[6] *Ibid*.
[7] "AI experts and public figures express their concern about AI risk," Statement on AI Risk, Center for AI Safety, accessed Dec. 16, 2023, https://www.safe.ai/statement-on-ai-risk#sign.
[8] *Ibid*.
[9] Scott Alexander, "Prediction Market FAQ," *Astral Codex Ten*, December 19, 2022, https://www.astralcodexten.com/p/prediction-market-faq.

advantage of the market to make money from inaccuracies in the pricing of shares.[10] This notion is supported by a prediction market asking whether Arnold Schwarzenegger would be elected as governor of California in 2003.[11] The two different markets tend to stay close in value avoiding a scenario where someone invests in one which is inaccurate, thus taking advantage of one market falling behind another. While in favor of the use of prediction markets, Alexander does note that these markets may suffer in accuracy when there is not a monetary incentive in place.[12] Also important to my analysis is the fact that prediction markets can be a useful tool in discerning the best guess of whether an event will occur or not when several experts disagree.[13] As seen by the context of the debate on the dangers of AI, this is the state that experts on AI find themselves in. The purported accuracy and state of debate on AI safety illustrates why prediction markets may be a powerful way to inform our views on the danger of AI.

There are other reasons why using predictions markets prove effective in examining the risk of AI. As Arrow et al. note, part of what makes prediction markets effective is the ability to aggregate information from diverse locations.[14] In the case of AI prediction markets, this means getting input from those that are knowledgeable about AI, but also those who may be knowledgeable about other areas important to the question of AI safety. For example, the market might receive input from someone like a historian who has studied other existential threats. They can make a prediction based on their understanding of how similar AI risk is to other risks

---

[10] *Ibid*.
[11] Justin Wolfers and Eric Zitzewitz, "Prediction Markets," *The Journal of Economic Perspectives* 18, no. 2 (2004): 116, https://doi.org/10.1257/0895330041371321.
[12] Alexander, "Prediction Market FAQ."
[13] *Ibid*.
[14] Kenneth J. Arrow et al., "Economics: The Promise of Prediction Markets," *Science (America Association for the Advancement of Science)* 320, no. 5878 (2008): 877, https://doi.org/10.1126/science.1157679.

like the risk of nuclear war. This gives the benefit of more information than just the input of AI experts who may experience some sort of bias within occupations.

Section III: The Danger of AI

It is clear that there are varying beliefs as to how dangerous AI truly is among experts. I proffer that an analysis of prediction markets–for reasons illustrated in Section I–can give way to a better understanding of whether or not AI poses a real threat to humanity. I will set out the following argument in this section:

1. If prediction markets demonstrate an equal or relatively larger chance of AI wiping out humanity compared to other existential threats, then AI should be perceived as a danger to humanity which needs to be acted upon.
2. Prediction markets demonstrate an equal or relatively larger chance of AI wiping out humanity compared to other existential threats.

   C: AI should be perceived as a danger to humanity which needs to be acted upon.

There is an important point which must be made about this argument. I specify that AI should be perceived as a danger–rather than it being a danger–purposefully. I do not wish to argue in any sense that the markets entail that AI is in fact a danger to humanity now or in the near future. What I mean to show with this argument and analysis is that due to the accuracy of predictive markets, one should follow the belief that the market prescribes as it is a best guess of what is true.

Regarding the first premise of my argument, I believe it is relatively uncontroversial. The specific existential threats I am referring to include a major nuclear war leading to the death

and inhabitability of Earth as well as a climate crisis reaching the point where Earth is no longer inhabitable. It is clear in the case of each disaster mentioned that humans perceive these threats as dangerous based upon their actions. Humans have created several treaties to reduce the number of nuclear weapons in the world, and many nations have refrained from using these weapons on another nation after seeing the devastation caused to the cities of Nagasaki and Hiroshima at the end of World War Two. Countries have also attempted to negotiate treaties to curb the carbon emissions of nations around the globe, and even individuals and companies have attempted to reduce their own emissions in recent years. From these actions, it is clear that there is an observed threat of climate change and nuclear annihilation which humanity collectively holds. So, if prediction markets which are relatively accurate rate the threat of AI wiping out humanity equal to or above the threats of these other examples of existential threats, it should follow that AI is perceived with the same or greater danger that these other threats are perceived with. From this perceived danger, AI research should be treated just as humans have treated the threats of nuclear war and climate change. There must be some sort of pause placed on advanced AI research to reduce the risk that AI poses.

An analysis of prediction markets illustrates the truth of the second premise. Looking at Manifold Markets, an online prediction market platform, there are several markets which ask predictors to rate the possibility of existential threats wiping out humanity in certain time frames. In regards to AI wiping out humanity, there are several markets on Manifold. I will focus on a market rating AI with a four percent chance of wiping out humanity by 2030,[15] a market rating

[15] "Will AI wipe out humanity before the year 2030?," Browse, Manifold, last modified November 17, 2023, https://manifold.markets/MartinRandall/will-ai-wipe-out-humanity-before-th-d8733b2114a8

AI with a nine percent chance of wiping out humanity by 2040,[16] and a market rating AI with a 16 percent chance of wiping out humanity by 2100.[17] I chose these markets to see a range of probabilities that AI wipes out humanity over a large time frame. I also utilize a market which rates climate change with a five percent chance of wiping out humanity by 2100,[18] and a market giving nuclear war a four percent chance of wiping out humanity by 2100[19] to compare these other existential threats to that of AI. The comparison of markets resolving in 2100 illustrates that AI has a roughly three to four times greater perceived chance of wiping out humanity by 2100 than two other commonly viewed dangers. Additionally, the trend of the chance increasing as the year increases across the three AI markets informs the view that the danger posed by AI is growing. Even if the chance of AI wiping out humanity in the next seven years is only four percent, there is a clear perceived potential danger looming around the corner if left unchecked by any regulation on AI. From the comparison of AI prediction markets and the markets on other existential threats, as well as the clear increase in chance of AI wiping out humanity as time goes on, it become apparent that an analysis of prediction markets shows that AI poses a similar or greater danger to humanity than other existential threats such as nuclear war and climate change. From this, and the acceptance of the first premise of my argument, it follows that AI should be perceived as a danger to humanity and a pause should be instituted on advanced AI research.

---

[16] "Will AI wipe out humanity before the year 2040?," Browse, Manifold, last modified November 17, 2023, https://manifold.markets/Tripping/will-ai-wipe-out-humanity-before-th-6d1e8126d974.

[17] "Will AI wipe out humanity before the year 2100?," Browse, Manifold, last modified November 17, 2023, https://manifold.markets/JamesDillard/will-ai-wipe-out-humanity-before-th.

[18] "Will climate change wipe out humanity by 2100?," Browse, Manifold, last modified October 16, 2023, https://manifold.markets/JosephNoonan/will-climate-change-wipe-out-humani.

[19] "Will a nuclear war wipe out humanity by 2100?," Browse, Manifold, last modified October 16, 2023, https://manifold.markets/JosephNoonan/will-a-nuclear-war-wipe-out-humanit.

A potentially common counter to my argument for perceiving AI as dangerous to humanity is that the markets I analyzed are not using real money like other prediction markets and may therefore be less accurate than a market which uses real money. It is true that Manifold uses a fake currency known as mana which cannot be redeemed for real money and can only be used to donate to charity.[20] As Scott Alexander notes, " if you can't make much money on a market, it won't necessarily be very accurate."[21] So, since Manifold does not allow individuals to make money off of an inaccurate market through the use of an artificial currency, those on Manifold may be less incentivized to correct a market which they believe is assigning an inaccurate chance of an event occurring. In the case of the AI markets, this means the chance of AI wiping out humanity could be skewed to be higher by some predictors who are overconfident in the certainty of such a scenario and a lack of other predictors taking advantage of an inaccurate market.

While the idea of making a profit is surely a strong incentive to make a market more accurate, I do not believe that this is the only way a market can be incentivized to remain accurate. It is noted in an article on the accuracy of prediction markets that predictors can be rewarded with social status rather than money earned through markets.[22] Based on additional reward systems, I argue that social status can be an incentive for predictors to make markets accurate, similar to how use of real currency incentivizes accurate action by predictors. Manifold meets this goal of incentivizing accurate markets by use of social status. Manifold's website includes a leaderboard which shows the top predictors based on their profit of mana

---

[20] "What is mana (Ṁ)?," About, Manifold, last modified 2023, https://manifold.markets/about.
[21] Alexander, "Prediction Market FAQ."
[22] James Tapper, "How Accurate are Prediction Markets?," *JSTOR Daily*, July 22, 2020, https://daily.jstor.org/how-accurate-are-prediction-markets/.

from their predictions.[23] It is reasonable to assume that predictors will understand that anyone can view the leaderboard and will view those on the leaderboard as more skilled than those with less profits. This understanding will lead to motivation to make enough of a profit to reach the leaderboard. To do so, one must avoid trying to manipulate markets away from what one actually believes, and one must also think critically about what they believe so they can identify when they can make a profit off of an inaccurate market. This description of how one should act to receive social status on the platform motivates the argument that social status incentivizes predictors to make markets accurate even in lieu of a monetary incentive.

Section IV: Conclusion

In sum, there are very different positions on whether AI poses an existential threat to humanity. While some assert that there is at least some danger, others discount the alignment problem entirely. By using predictive markets, I demonstrate that the perceived danger is at the very least as strong as the perceived danger of other existential threats of climate change or nuclear war. From this equation of danger, I argue that humanity must act to lessen the danger through a pause in further advanced AI research.

There is much more to be covered in the debate over AI safety in the future. More work can be done in discussing what sort of pause is best for humanity to implement in the near future. Different tools such as the Total View in population ethics or Quality-Adjusted Life Years could be used to make more sense of what is and is not ethically required of AI scientists. It may also be helpful in the debate to create more prediction markets which use real money to ask about whether AI or other events could wipe out humanity in the future. While I stand by the use of Manifold markets even though the site uses artificial currency, gaining the ability to analyze

[23] "Top Traders," Leaderboard, Manifold, last modified November 17, 2023, https://manifold.markets/leaderboards.

more markets which use real money can only help to shed light on the true danger of AI and other possible existential threats.

Bibliography

Alexander, Scott. “Prediction Market FAQ.” *Astral Codex Ten*, December 19, 2022. https://www.astralcodexten.com/p/prediction-market-faq.

Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, et al. “Economics: The Promise of Prediction Markets.” *Science (American Association for the Advancement of Science)* 320, no. 5878 (2008): 877–78. https://doi.org/10.1126/science.1157679.

Center for AI Safety. “AI experts and public figures express their concern about AI risk.” Statement on AI Risk. Accessed Dec. 16, 2023, https://www.safe.ai/statement-on-ai-risk#sign.

Manifold. “Top Traders.” Leaderboard. Last modified November 17, 2023. https://manifold.markets/leaderboards.

Manifold. “What is mana (Ṁ)?.” About. Last modified 2023. https://manifold.markets/about.

Manifold. “Will AI wipe out humanity before the year 2030?.” Browse. Last modified November 17, 2023. https://manifold.markets/MartinRandall/will-ai-wipe-out-humanity-before-th-d8733b2114a8.

Manifold. “Will AI wipe out humanity before the year 2040?.” Browse. Last modified November 17, 2023. https://manifold.markets/Tripping/will-ai-wipe-out-humanity-before-th-6d1e8126d974.

Manifold. “Will AI wipe out humanity before the year 2100?.” Browse. Last modified November 17, 2023. https://manifold.markets/JamesDillard/will-ai-wipe-out-humanity-before-th.

Manifold. “Will a nuclear war wipe out humanity by 2100?.” Browse, Last modified October 16, 2023. https://manifold.markets/JosephNoonan/will-a-nuclear-war-wipe-out-humanit.

Manifold. “Will climate change wipe out humanity by 2100?.” Browse. Last modified October 16, 2023. https://manifold.markets/JosephNoonan/will-climate-change-wipe-out-humani.

Piper, Kelsey. “A Field Guide to AI Safety.” *Asterisk Magazine*, June, 2023. https://asteriskmag.com/issues/03/a-field-guide-to-ai-safety#the-intelligence-explosion-and-early-work-on-preventing-ai-catastrophe.

Tapper, James. “How Accurate are Prediction Markets?.” *JSTOR Daily*, July 22, 2020. https://daily.jstor.org/how-accurate-are-prediction-markets/.

Wolfers, Justin, and Eric Zitzewitz. “Prediction Markets.” *The Journal of Economic Perspectives* 18, no. 2 (2004): 107–26. https://doi.org/10.1257/0895330041371321.

Yudkowsky, Eliezer. “Pausing AI Developments Isn't Enough. We Need to Shut it All Down.” *Time Magazine*, March 29, 2023. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.