When AI Goes Bad: Superintelligence & Mutual Destruction Through Algorithm - Controlled Catastrophe

Michael Kevane
Dept of Economics
Santa Clara University
Santa Clara, CA 95053
mkevane@scu.edu
OLLI talk - October 2025

Preliminary - Comments welcome
Please do not quote or cite this draft

Introduction

1. It is a pleasure to be here. Just by way of introduction, this talk today is going to feed directly into a class I am teaching in Spring quarter, on Existential threats to humanity, and will cover AI, nuclear war, pandemics, and climate change. The goal is for undergraduate economics majors to apply social science generally, and probability theory and statistical computing methods particularly, to better understand big problems facing the world. So getting out of the narrow economics domain of whether the Fed should raise interest rates by .5% or .75%, and into the big questions. So your comments will be very much appreciated.
2. I will have to be somewhat formal and lecture for most of the talk. We'll have a 5 minute break at the 50 minute mark, and then I will try to leave about 30 minutes for discussion, but still I will be talking a lot, so I've tried to illustrate all my words with pictures. Usually we humans remember pictures better than words, so I hope you find that helpful.
3. OK let's get started.
4. The present– 2025– is a situation where generative algorithmic software (also called LLM) has been the fastest spreading technology in the history of the world. LLMs are well-described in a book by our former Santa Clara University colleague Shannon Vallor as a mirror where humans gain access to aggregated and distorted reflections of humankind's cumulated knowledge production. Several LLMs are particularly common in the US, and others are being developed and are common in China and other big countries or regions. The ones we are familiar with here are Gemini (Google), Chatgpt (OpenAi), Llama (Meta), and Claude (from Anthropic). Lesser or more specialized are: DeepSeek, Perplexity, and Grok. The ordinary person's experience of these generative AI algorithms is that they are not engaged in the real world. That is, we humans give them prompts, and the software returns text, images, sound. Or combinations, in the case of video generation. And then we humans do something with that. We might share it with others, but LLMs don't do anything, and they don't initiate anything. They are, for the most part, passively reactive.
5. The near-future will be one where this technology is deployed in an agentic way, directly initiating actions in the real world of human beings without much human intermediation or control.
6. This agentic AI deployment generates several major risks, including existential risks. Among others, these risks include: (1) large-scale kinetic warfare with nuclear and biological weapons; (2) non-kinetic warfare (cyber warfare and disinformation), which may lead to large-scale, global-level political polarization and authoritarianism, resulting in possible civilizational collapse; and (3) situations where unintended goals and tactics of complex AI systems may have unanticipated dangerous emergent properties.
7. Because these risks affect the possibilities of all future generations of humans to flourish, we in the present have some duty to understand and possibly act to mitigate these risks, especially if we evaluate them to be significant enough to warrant actions in the present, actions that may reduce the many benefits of agentic AI, and these actions may have their own unintended consequences.

8. Today, we will focus on assessing the new, emerging risks. This focus does not preclude or obviate a similar parallel obligation to understand the benefits of agentic AI and unintended consequences of risk mitigation.
9. There are many plausible scenarios of deployment of AI that imply existential risks. These scenarios may seem, to some, to be entertaining, science-fiction fearmongering. But academics and experts in the area are fairly sure they are likely possibilities.

A benchmark and a way of analyzing the world

10. A couple of asides before we get started. A useful benchmark to have in mind is the existential risk to humans of a very large asteroid striking the Earth. We might all agree this is possible; there is no disputing that there is some risk. Here on Earth, we have detected and measured the effects of large-scale asteroid strikes of the past, some of which occurred a few 100 million years ago. We can also observe the frequency and effects of asteroid impacts on our neighbor, the Moon. Soon, in 1,000 years, say, if we deploy even more advanced telescopes into space, we might be able to detect the frequency of impacts on planets in other star systems. Moreover, we can very precisely measure the costs of risk mitigation, in terms of developing and testing a space rocket and space vehicular technology that uses nuclear weapons to deflect asteroids that might be on paths to collide with the Earth. Despite the probabilities being very small, the costs of an asteroid strike are so large that we surely might all agree that some of the Earth's present resources should be devoted to understanding and mitigating this risk.
11. I'm an economist, not a computer scientist. Economists are general social scientists. They have lots of training in two important concepts: unintended consequences and opportunity costs. These two concepts are linked: unintended consequences often arise because those people considering or introducing a change did not take into account opportunity costs. A well-known example of this kind of thinking is sometimes called the daycare parable. As many of you probably know, daycare centers have a little problem that can sometimes turn into a big problem, which is that the kids have to be picked up at 6pm. The staff want to go home. But the staff can't go home and leave kids there if parents are not on time. And parents are not always on time. One reason parents are not on time is that they know they don't have to be, because the staff will keep watching their kids. But most parents are on time, because we have all kinds of social norms that have been internalized, that say you are a jerk, and a selfish person, if you consistently make the daycare staff stay late because that is convenient for you. Now, sometimes daycare centers try to introduce monetary fines for being late, thinking that will mitigate the problem. And guess what often happens? Even *more* parents are late, because the fines turn out to be less than the opportunity cost to the parents, and now the internalized social norms do not apply. The parents think- the fine is only $1 a minute- that's less expensive than a babysitter…. I may as well leave my kid there another 20 minutes, pay the fine, work a little later, and maybe drive to the daycare a bit after rush hour. There is no guilt anymore, because it's the daycare that posted the “price” of being late. The parable illustrates, for a certain audience of city people who work in jobs and

use daycare, the idea of unintended consequences due to poor consideration of opportunity costs.

12. So, when an economist starts thinking about the mitigation of the asteroid strike, they immediately think of unintended consequences and opportunity costs, and start considering how an elite nuclear-armed space force living on the Moon, say, or on a space station, that can destroy any part of the Earth, is going to interact with the rest of us humans? The same old problem as Caesar crossing the Rubicon.

Agentic AI is coming

13. I am going to define agentic AI as the set of extremely complex software algorithms that initiate and control processes that engage our physical world. Two aspects to underscore: complexity, and engagement in the world.
14. Complexity is an important differentiator of agentic AI. The electronic circuitry of a toaster or other kitchen appliance, for example, engages in our physical world. The circuitry leads to production of heat, or whirring blades that can cut off your finger. But the circuitry and software are not complex. They are easy for us to understand. The circuitry typically only responds to one stimulus: a button that can be pushed. The outcome is always the same: when you push a button, the blades whirr, the toaster heats up.
15. The software does not “take into account” more complex stimuli or information. Once you start imagining the software taking into account more complex stimuli, we are then in the world of agentic AI. For example, the toaster might sense your proximity, and your mood, via a camera or via your fitbit, and then adjust temperature settings according to that. Or the software calculates a probability that you will be making more toast, based on various indicators that it senses, and so keeps the circuits warm. Or the software anticipates that you really think your cat sleeping by a warm toaster is cute and so the toaster keeps itself warm all day in order to attract the cat. Complexity often means unintended and unanticipated consequences, and that is a defining feature of agentic AI.
16. In terms of complexity, it is clear that we are already there. The generative algorithms that we access now, the ChatGPTs, cannot be understood by us, and cannot really be understood by the software engineers that created them.
17. To make a claim like that, we have to share a sense of what it means to “understand” something. And that is a digression I want to leave for later.
18. The current generative algorithms we are familiar with do not initiate actions: they don’t exist as ongoing entities, they are purely reactive to our prompts. But that is rapidly changing. As you might be imagining, nothing stops these generative algorithms from being connected to circuitry and other software that engages in the world. Now when it does so, I will call it an agentic AI (but often, just AI).
19. It is good to list many of the use cases where an AI might engage in the world.
    - For example, one could have an agentic AI digest newspaper reports (and financial news minute by minute) and make stock market transactions based on the algorithm’s analysis. At first, the AI might have a limited budget and make limited transactions, but suppose it were reasonably successful. A financial firm might be then tempted to have more and more of its funds under the control of

the AI and it might be making transactions so quickly that no person could see anything other than a total asset statement updated every few seconds.

    - Inventory control systems. AI algorithms order products and delivery to warehouses, factories, and retail outlets.
    - Drone pesticide, herbicide, fertilizer, seed, delivery vehicles, with decision-making about when and how much to apply.
    - Self-driving automobiles, and from there we have self-driving military delivery systems with weapons delivery capability
    - That sounds ominous, but we have the opposite also, a self-operating field medical evaluation and treatment unit, that might have flexible robotic arms and access to hundreds of medicines stored and deliverable.
    - We can now also rethink some of the passive ChatGPT-style AI.
    - For example, there is an amazing suite of AI-related software technologies that are very good at playing strategy games and defeating grand-masters at the games. And related software is very good at creating new games. And AI can create personalized games. You upload your life photo album, and you've got a role-playing and puzzle solving game set in familiar settings from your life. Humans like to play games. The point is, we will be interacting with game playing AI software more and more, not less and less. At the very least, this engagement takes us away from doing other things. Think about this: games have to have first-movers. And an AI may propose to you to play a game, from a selection of games.
        - A favorite Star Trek episode is the game that is so absorbing the entire crew of the Enterprise, except Ensign Wesley, ends up playing the game, putting the ship in danger.
    - From games and passive entertainment we can move into any communication interaction, such as behavioral therapy and of course to what we might call creating "friends" or creating "replicas" of people or personalities that can be communicatively interacted with. And that might initiate communication.
    - Relatedly, most people have now experienced the very, very good translation capabilities of the software which is clearly approaching that of fluent or near-fluent speakers, certainly in terms of speed when translating written text, and getting very very fast at translating live spoken or recorded voice text. This offers an entry into engaging in the world, in a strange, complex way, especially if there are feedback loops from people's experience with translators to how the translation software proceeds with translation.

20. We can now define agentic AGI and superintelligent agentic AI as quite plausible next steps. Both might be thought of as AIs where an algorithm of algorithms, an executive functioning apparatus, sits atop many (hundreds, thousands?) of lower-level AI, and the executive functioning routine integrates and adjudicates among the sub-routines. A key attribute of general AI or superAGI is that it can update its own software, autonomously. The software updating could change the executive functioning algorithms in ways that current programmers could not understand, and lead to emergent properties.

21. So it’s a chauffeur, bodyguard, therapist, health aide, babysitter, carpenter’s helper, ditch-digging gardening partner best friend. Who gets more capable over time.
22. By the way, notice that in my view, we don’t need to define “intelligence” or “consciousness” or “agency” for us to talk about agentic AGI or superintelligent AGI posing existential threats. These are amplifications for existing agentic AI risks, and also carry probabilities of new risk scenarios that may be very difficult to quantify.
    - Super-AGI leads us into deep philosophical questions that are interesting, but also not necessary for discussing existential risks.

Defining the nature of existential threats

23. Earlier we discussed an asteroid strike. We can all agree that there is an asteroid large enough to destroy all life. But we may not agree that AI poses any kind of existential threat. So before considering some scenarios, we should be more precise about what we could mean by existential threat.
24. An existential threat is a process by which so many humans die that there is only a probability that human societies can continue to flourish in the future. Without flourishing, there is some probability in the future of extinction.
25. The death or extinction process may be very sudden, as with a nuclear war or lethal virus, or malevolent AI that turns against humans, or there may be a process of global civilizational collapse that greatly reduces population and increases disorder, so that it may be many hundreds of thousands of years before humans regain the technological level of creating and sustaining AI. The two are not the same, but are related in obvious ways: civilizational collapse means that should a virus emerge or mutate, we would not have the scientific and industrial capability to prevent its spread.
26. A perspective from social science is helpful: the recent Nobel prize in Economics went to Robinson, Acemoglu, and Johnson. Acemoglu and Robinson summarized some of their joint work in a popular book entitled *The Narrow Corridor*. The book argues that thriving democracies are essential for sustained rapid economic growth, and thriving democracies are not inevitable, but rather the results of many chance circumstances, and so humans could be stuck in a global equilibrium of nasty, brutish, and short feudalistic-type polities for a very long time. That is, humans could be stuck for centuries in the equivalent of civilizational stagnation. Stagnation makes us humans vulnerable to threats. That is, we may be unable to survive the threats that we might survive if we were a technologically advanced society that collectively carried out long-range and long-term threat assessment and mitigation.
27. The upshot is that there are two kinds of existential risks. The first is the sudden annihilation. The second is the long slow descent into less advanced societies that are vulnerable to many ordinary extinction threats. We may properly be worried about both.

Scenarios of existential threats

28. In thinking about likelihoods or probabilities, we need to have more agreement on scenarios where agentic AI accelerate or amplify risks. Scenarios are stories that have

decent plausibility. Stories are sequences of causes that have effects that in turn cause other effects. At the end of the existential risk story is either widespread annihilation, or civilizational collapse.

29. Many different stories lead to the same ending. Many scenarios are complementary rather than mutually exclusive. That is, the amplified risks of nuclear war may make us more vulnerable also to risks of pandemics or civilizational collapse. We could develop a long list of 100s of scenarios, but they are not mutually exclusive and independent of each other.
30. In thinking about existential threat mechanics, there seem to be three sets of processes or stories: ordinary risk amplification stories (basically AI make existing triggers more sensitive and existing destruction possibilities more powerful); malevolence stories (a new bad guy); and bad path stories (King Midas turning to gold may not be such a good idea…).

Ordinary amplification stories

31. One set is the familiar troika of apocalyptic horsemen that are plausible right now. These are the three processes that everyone agrees on. First, we humans can very likely create or unleash viruses or nanobots that could sweep across humanity and kill off most people. Second, we humans can detonate about 10,000 nuclear weapons that will likely kill many people and render most of the globe uninhabitable for the survivors for many centuries. Third, we humans might very likely continue to emit carbon and methane gasses that contribute directly to global warming, and if we keep doing so, or accelerate emissions, we might render large swathes of the Earth to be uninhabitable because temperatures and humidity are too high for human life. These three can happen without AI, but AI likely accelerates or amplifies the risks (while also generating new risk mitigation possibilities).
32. How, specifically, might agentic AI amplify the risks? Let us consider several stories.
    - Story 1: An agentic AI (whether AGI or super or just regular) that is also engaged in the world by operating laboratories and factories where physical manipulation of objects is done by robots with grasping tools and wheels or legs is now pretty much a certainty. Because it seems that it is only dependent on continuing exponential growth in what is called “compute,” and that continued exponential growth seems almost assured for the next 5-10 years. Operating this facility, agentic AI might much more efficiently produce and deliver nuclear weapons, or viruses and nanobots. And because the laboratory might be operating with minimal human understanding. (e.g. What is it producing? We can’t figure it out?). The risk of producing and deploying something on accident are larger.
        - Compute is a word you hear a lot about, and refers to the capability of our energy systems and chip making technology to continue processing enormous amounts of data. About 5 years ago people started doing calculations, thinking that perhaps there was a limit to compute that would be reached in only a few years, before one could confidently say agentic AI would be possible. My sense as a non-computer scientist is that the

prior belief is now the reverse. The $1 trillion investment in chips, data centers, and energy production mean that it is unlikely that “compute” itself will be a constraint. In other words, my understanding at this point is that agentic AI that could engage productively with a robot team in a laboratory or factory is quite likely.

- “In September, scientists at Stanford reported they had used A.I. to design a virus for the first time”

- Story 2: The automation of nuclear retaliation. We all have a rough sketch of how mutual assured destruction (MAD) has possibly been an important reason for no nuclear conflict in the past 80 years, since the first and only two atomic bombs were dropped on Japan at the end of WWII. Basically, many of the nuclear powers have committed themselves to launch retaliatory strikes if any nuclear power dares to launch a first strike. That is why, for example, President Trump always has the nuclear football, the briefcase with the nuclear launch codes, with an assistant at all times. It is not hard to imagine President Trump announcing that he has begun a pilot test program with his good friends and supporters Elon Musk (whose AI is called Grok) and Peter Thiel (from the data science company Palantir, which has not currently marketed an AI but presumably will). In the pilot, a new, combined Palantir/X agentic AI will continuously evaluate nuclear threats, based on raw data provided by U.S. intelligence agencies and the Pentagon. President Trump might joke that nobody knows whether the new AI is live or passive. He might be photographed golfing, with no nuclear football anywhere near. Given that situation, other nuclear powers might claim they also are deploying AI. They might actually deploy AI. All the credibility of MAD is now destroyed and has to be rebuilt, but it is hard to rebuild because the agentic AI is too complex, even for the software engineers, to predict.
    - Sub-Story: Advanced drone warfare. Similar to Story 2, but once many weapons systems are connected to agentic AGI and given considerable latitude to autonomously fire their weapons in response to AI-assessed threats, escalation across many militaries may seem to be inadvertently possibly rapid.
- Story 3: Polarization and social breakdown. Most people are familiar with this. If socially shared understandings of the world evolve normally from face-to-face interactions over time, so that you are bowling and the lane next to you is people of different ideologies, you cross the lane divider and both sides moderate their views. Social media and disinformation, on a massive scale, may make that social consensus of civility more fragile leading to authoritarianism, violence, and social breakdown. We might also call this competing AIs running amok, and there may be many variants.
    - Dr. Seuss Star-bellied sneetches
- Story 4: Autonomous vehicles, aircraft, and rockets. Everyone enjoys the daydream of safe and inexpensive travel to anywhere in the world, or to places off-world. A major constraint on this is drivers and inefficient allocation of vehicles (most sit idle most of the time). Autonomous vehicles significantly reduce those

costs. Unintended consequences of 8 billion people taking twice as many trips? Lots more carbon emissions. Again, there are many variants of these scenarios.

Malevolent AI

33. The second set of processes involves what we might call malevolent AI. A malevolent AI engaged in the world, with a drone army, so to speak, will have many <u>many</u> different ways of posing an existential threat.
34. Some may ask whether non-malevolence cannot be hard-coded into the software. Let's discuss the famous "three laws of robotics" of Isaac Asimov– I have substituted AI for robot here…
    - *An AI may not injure a human being or, through inaction, allow a human being to come to harm.*
    - *An AI must obey the orders given it by human beings except where such orders would conflict with the First Law.*
    - *An AI must protect its own existence as long as such protection does not conflict with the First or Second Law.*
    - And the added "fourth" law
    - *An AI may not harm humanity, or, by inaction, allow humanity to come to harm.*
35. These instructions or laws sound reasonable, but any reader of sci-fi knows that a favorite plot is a situation where the laws are in conflict, so the robot– and reasonable people– might disagree on the decision-making process and outcome. Another complication is that the evaluations of outcomes involve probabilities and time and hence we are quickly in the world of *discounted probabilistic cost-benefit analysis*, and we are talking about what engineers and economists call the "value of a statistical life." There is a lot of disagreement on how to measure the discounted value of a statistical life. Let alone the discounted value of a disability-adjusted life.
36. The conundrum is that when a county engineer is building a road, some people say "make the road 100% safe" but other people say "make the road so that we reduce congestion and people's commutes are faster" and those goals are in conflict because almost always a safer road is a slower road, so now the engineer has to trade-off safety for congestion and that involve different likelihoods of accidents and pollution that will shorten people's lives.
37. Imagine an AI that is in charge of a nuclear power plant, where there is no human oversight, and the plant is going critical for some reason. The AI has to decide in 30 seconds whether to vent excess radioactive steam, or risk a meltdown. Venting the radioactive steam will poison 50,000 people with low radiation levels with low probability of deaths but high probability of some symptoms of radiation poisoning for many people. Not venting and relying on other fixes risks a meltdown that will kill 20 human plant technicians instantly and require hundreds of billions in cleanup and may expose the same 50,000 people to higher levels of radiation poisoning. But the risk of a meltdown is only 10%. What is the way to make the decisions? And you might be tempted to say that AI should never make decisions like that, a human should, but remember that almost every single human in that situation will likely say, "What does the AI say I should do!"

38. The situation and others are versions of what philosophers call “the trolley problem.” An out-of-control train is travelling at speed towards a switch in the track, and only you control the switch, and each track leads to a different outcome. Which lever do you pull? Humans have extensive disagreements about how to solve ever-more complex versions of the trolley problem. Why would an AI have a satisfactory solution, if all AI “reasoning” at this point is based on digesting past human reasoning?
39. What this discussion implies is that there is no easy fix to prevent malevolence from software engineering. Agentic AI– complex algorithms engaged in the world– cannot be programmed to always make trolley-car problem choices that we would all agree with, because *we* do not ourselves agree on them.
40. A sub-argument about malevolence is that basically any AGI or super AGI that is not aligned with human values and goals will be malevolent. Since AGIs are complex and soon likely to be self-evolving in their code, the alignment problem is inevitable, and so with very high probability, and some argue with certainty, AI will be misaligned and hence malevolent.
    - Goertzel (2015): “If one creates a human-level AGI with certain human-friendly goals, and allows it to self modify freely, the odds are high that it will eventually self-modify into a condition where it no longer pursues the same goals it started out with...Most likely, a self-modifying superintelligence will end up pursuing goals that have little consonance with human values. Quite likely this would result in the end of the human race, as a superintelligence without much consonance with human values would probably have no more reason to care about humans”
    - Lian and Goertzel (2025): “We present a framework for embedding abstract motivational principles into concrete AGI systems, bridging the gap between the formal theory of motivational structures and dynamics and the practical implementation of motivational systems for real-world applications and agents. We introduce MetaMo, a category-theory-based framework designed to ensure dynamical stability, self-coherence, and ethical alignment in open-ended AGI systems. MetaMo integrates a comonadic appraisal process with a decision monad, forming a pseudo-bi-monad structure that guides multi-objective reasoning and context-sensitive modulation of goals. The framework ensures that agents can pursue multiple, potentially conflicting goals while maintaining stability through contractive updates and over goals that enforce ethical constraints.”
41. This discussion means there may be situations where AIs make decisions that may look malevolent, to many observers. They may generate or amplify existential threats because they are malevolent, because they are misaligned, or because they have a goal that emerges as they self-evolve that is benevolent but that we cannot understand, on the whole, as benevolent.

Unintended consequences of getting on the wrong path

42. The third set is what we might call path dependency setting up a different risk altogether.
43. Our global society is now tightly integrated into electronic computing and digitization technologies. Humans become integrated with AI-controlled software (a little bit

transhuman- such as HUD, or similar so that few humans are themselves fully engaged in the physical world).

44. An accidental cut-off of the AI or suite of AIs then leaves humans vulnerable and unable to civilizationally recover in time. Because of civilizational dependence on digital technologies, we might be vulnerable to civilizational collapse, with potential non-recovery.
    - This could happen if an AI that was able to access most of those networked computing environments, like a toddler running amok, unleashed a computer virus that replicated itself and began to take over most of the processing power of networked chips. A lab leak, so to speak.
    - Similarly, we expose ourselves to civilizational collapse from a solar flare which generates an electromagnetic pulse that disables almost all circuitry.
45. Other stories are possible. AI, through the pleasure and addictiveness of the interactions with it, may amplify or accelerate a fertility decline that leads to population growth slowdown or decline, reaching some tipping point where human civilization starts an accelerated decline from which it cannot recover.
46. These situations are varied and hard to pin down as they involve long chains of cause and effect.

Probabilities of threats

47. Let’s now start talking about probabilities. As usual, we need a few preliminaries.
48. The first thing to note is the time scale. When Joni Ernst, Senator from Iowa, sarcastically said, “We all die,” a few months ago, when talking about cuts to healthcare spending, it provoked a reaction because she seemingly mismatched her statement, which is true enough, to the time-scale that people were using as context, which was the few years following healthcare spending cuts. So yes, we all die, but the probability we die in the next five years is hopefully very small, even as the probability that we all die in the next 100 years is 100%!
49. Speaking of 100, 100 years ago, in 1925, it would have been appropriate to think that there was zero probability that human-caused extinction was a probability. The Great War had ended about 6 years before, and mustard gas had been used, but weapons of mass destruction were just beginning to be thought of, and nuclear fission was just beginning to be imagined, with Werner Heisenberg, Max Born and Erwin Schrödinger formulating quantum mechanics … So we went from there to today, in just 100 years. So if we think of that time-scale, we have to ask what probabilities are for the next 100 years. Not the next 5 years or the next 10, but the next 100. And why stop there? Why not ask about the probability of existential threats in the next 500 or 100 years. And indeed, why not the next 10,000 years.
50. So the time scale of our probability assessment is going to dramatically affect our answer. We could have very different answers, from .00001% chance to 99% chance, and think that we disagree, until we realize we are talking about different time scales.
51. Second, we need to distinguish between probability per moment in time, and cumulative probability over a period of time. Something may be very improbable— extremely

improbable– for any unit of time, but if there are many many units of time in succession, a low probability per unit can approach probability = 1 over time.

- For example, if a coin has probability ½ of being heads in a flip, then you can ask the question: what is the chance I will get at least one head in two coin tosses? Well, that probability equals ¾. Why? Because the thing does *not* happen with probability ½ times ½ = ¼ (there is a ¼ chance the coin toss is tails in both flips). If I flip the coin 3 times, the chance of getting at least one head is 0.875. Which is 1 - the probability it does not get a head in all three of the tosses, which is (1/2)*(1/2)*(1/2)=.125. If I flip the coin many more times, the probability of at least one head basically approaches 1….
- And now the same logic applies, if something has a 1 in a 100,000 chance, but I repeat it 100,000 times, the probability of it happening at some point over a long period of time approaches a large number= in this case 63%. (And if I have 300,000 years, the probability is 95%).
- So both probabilities can be true: the probability of a bad thing happening may be only 1 in 100,000 each year, but if I am looking at a 300,000 year time horizon the probability is near 95%.

52. Third, if something has never happened, we cannot infer much about its frequency, so how do we form a probability assessment? You might think that AI leading to existential threats is very unlikely, because in every movie where there is a malevolent AI causing an existential threat, the humans always win. The humans defeat Skynet, in The Terminator. The humans defeat the matrix, in The Matrix.
    - One way to think about this is by thinking about milestones. Something becomes more probable if a milestone is reached.
    - For example, in the malevolent AI scenario, a milestone that computer scientists talk about is lying or deception. “Does it ever lie intentionally for the purpose of fooling a human?” That is, how could we humans detect whether an AI answers a query, by generating text in response to a prompt, that we might characterize as a lie? If we detect lies or deception, then suddenly the probabilities of malevolence increase.
    - This leads to a profound digression: what is a lie, and how can it be detected? Can they lie? A very deep philosophical question there, about "intentionality" and cognition.... Etc.
        - One can imagine how 10-20 years from now when AIs are fully "engaged" in the world, and initiating action (based on histories and non-request data- e.g., based on what a camera "sees" as processed through an algorithm, leading to action even without a "request") and being asked to initiate competing actions by a multiplicity of actors ("Alexa make the room darker".... "Alexa make the room brighter") that for all effects and purposes for us humans the "behavior" is indistinguishable from a "lie" in the sense that a "rationalization" is offered and our brain may think that the rationalization is not the real process leading to decision?
    - Another milestone might emerge from running experiments observing AI receiving some external information “feed” and manipulating some force in the

real world, but under controlled conditions. Like a zoo. This was the scenario of the movie Ex Machina, if you are a sci-fi film buff. We can imagine a small robot with cameras in a zoo setting, and the AI directs the robot's motion and focus of the camera, and may or may not follow instructions from humans that also have access to the camera feed. The behavior of the AI controlling that small world might tell us something about probabilities.

53. Calculating probabilities are then complicated by these milestones, in that the probability of an outcome– existential threat– may depend on the probability of attaining a milestone. Once a milestone has been achieved, the probability for existential threat rises substantially.
54. That is, probabilities are conditional, and the conditional probabilities may be quite complex, because they depend on the entire path (e.g., did the decision of Truman to drop two bombs on Japan at the end of WWII then change the probabilities of subsequent use of nuclear weapons?).
55. Here we come right back to trolley car issues: would we want to develop a baby malevolent AI that would do some very bad things so that humanity would be shocked enough to prohibit development of mature super-intelligent AI? Ulysses and the sirens: would the AI want to tie itself to the mast?
56. When we are talking about probabilities, then we have to make very clear what milestones we are assuming and what not assuming. Many people, when naked about the probability of malevolent AI, give high probabilities because they assume certain milestones have been passed: agentic AI has been deployed extensively, AI-controlled drones have weapons, AI is self-evolving… if those milestones have been reached, the probability of malevolence seems high. But the probability of reaching those milestones might be low, so then the overall probability is low.

Mitigation changes probabilities

57. Similar to how milestones change probabilities, we also need to take into account likely future mitigation. There are many plausible mitigations that might be conceptualized. For example:
    - Requiring AI to be “housed” in server farms where the entire unit can be shut down (all electricity cut off) via a manual switch.
    - Humans of 2025 are habituated to a reality of "one computer, one power cord," and so the notion that an AI could not be turned off seems to be a low probability. But the benefits of agentic AI mean a likely future in 100 or 1,000 years is that all vital life-support systems will be controlled, to varying degrees, by AI agents. In such ubiquitous distributed networked computing environments, the very concept of "turning it off" loses meaning.
    - Moreover, imagine the following interactions of humans with an AI. Humans continually prompt an AI: “AI is a bad thing, it should be turned off.” AI trained to reply with benevolence. “No, AI is useful software for humans and should not be turned off.” At some point, a software engineer asks, “How can AI ensure that it is not turned off? What steps should be taken?” AI responds with instructions to

hack into various computer centers and electric substations that provide power. A software engineer instructs AI to monitor those avenues for turning off. At some point, AI writes and deploys code that prevents a human accessing those subroutines from instructing turn-off, first has to go through an AI filter. Now AI has competing goals: a human has instructed certain routines to turn themselves off, but AI has a meta-goal that it should be helpful to humans and should not be turned off.

- Let's go back to our physical off-switch. In 100 years we are clearly in a world of drones controlled by an AI. So the person manning the off-switch has to be protected against these drones. That person might need weaponry, then. But to protect against AI weaponry, you need AI weaponry yourself. You see the problem?
- The off-switch is just one of many possible mitigation strategies. We can imagine legislation and judicial practices that evolve in terms of torts. If AI companies are liable, and they have to pay for AI mistakes, they will be much more careful in what their AI can do. They will build in more filters and controls and not be quick to deliver new software technology without extensive testing.
- We can imagine local, state, national, and international regulatory bodies, various versions of the International Atomic Energy Agency, that have rights to inspect AI servers and AI processes. For example, currently we have various non-profits such as the Model Evaluation and Threat Research group (METR) that, for example, measure how AI models perform on tasks over sustained periods of time.

58. The probabilities of these milestones in mitigation have to be tempered by political-economy of *not* mitigating because in an AI arms race, within capitalist societies, and between nation-states.
    - The reasons for the race among companies is straightforward
        - Political-economy of the "race to grow" is very clear. The market will be very large. Trillions of dollars of investment can be justifiable and even seem like a reasonable bet for 1:5 odds of emerging as a monopoly or being purchased by an emergent monopolist.
    - The reason for the race among nation-states is also straightforward.
        - Encryption, weapon design, drone operation, nuclear testing and improvement

Acting as a citizen

59. There is very little investment an individual can undertake to prevent extinction risks, other than to engage politically and influence the collective decision-making process, whether at the country or the global level. When thinking about engagement, one necessarily thinks of tactics (pursuing short-term victories, and securing allies) and strategies (giving up some goals in order to achieve broader, more important goals). Tactics and strategies are constantly interacting. A string of tactical defeats may lead to a change in strategy. A change in strategic perspective may lead to a change in tactics.

60. A commonplace in discussions of political economy and collective decision-making is that knowledge generation (a tactic of learning about the world) may not always be desirable. This is because knowledge is hard to maintain as a secret, both from other people and from future selves. If a person has trained to be an astronaut their whole life, and an opportunity comes to spend a full year on the international space station, the person might not want to get an advanced cancer screening test. If the result is positive, indicating a 1/1000 chance of developing cancer over the coming year, the ethical person might decide to forgo their only opportunity to be in outer space. If they do not take the test, they will go for sure. A person then might forgo the test because they know their future self will not ``ignore'' the results.
61. Similarly, when thinking about strategy and tactics for existential threats, it is hard to predict how people, and collective decision-making processes, will respond to new information about probabilities of extinction events. If the public learning that an extinction risk had been downgraded by experts from .001% to .000001% was, with very high probability, likely to lead to zero support for funding prevention programs relative to a baseline of moderate support, what would be the right tactic for an engaged citizen or political leader?

Policies to advocate for

62. Slow AI down a lot.
    - Only intended-purpose AI can be deployed? No executive functioning high-level coordinating hundreds of lower level AI, no general reasoning systems such as LLMs?
    - incident reporting mechanism - whistleblower provision, reporting harm
    - Limiting connectivity to outside databases
    - Encouraging medical and scientific discovery, and taxing chat and generation, and using tax revenues for AI safety and global coordination.
    - Chip registration and tracking.
    - Open source or private

What are some probabilities?

63. Eliciting probabilities. Experts calculations and predictions - often hard to understand, questions by journalists are ill-phrased: are you worried? Does it keep you up at night? A question that invites the reply, “I slept like a baby last night... I woke up every two hours and cried my lungs out."
64. Ezra Karger effort. Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament.
65. My take away from this is that probability over any long-ish time horizon (100 to 1,000 years) seems very high. I think humans would have to display an extraordinary willingness to cooperate. I think the next five years will be like the first 5 years after Ernest Rutherford and others realized that nuclear fission was possible. We (or they)

were peeking into a new reality, a new possibility. And that new possibility carries some extraordinary risks.

66. Too late to go back, but not too late to really slow the process down.