This article gathers key findings and arguments across three key areas: (1) AGI timelines; (2) Existential AI risk; and (3) Policy responses. The goal is to fairly portray the best arguments on each side in the first two sections and ultimately synthesize each section (with the author providing opinion commentary during this step). This is intended to be a living synthesis. Did I misunderstand or miss a key argument? Is my analysis weak? Does this need to take into account a recent breakthrough?

Intro

We may live in *the* most interesting time.

If GDP growth continued for 8,200 years at 2%, all atoms in the galaxy would have multiple economies bigger than the current world economy. Considering the universe is billions of years old, we're living in a top 0.00001% period. Our rate of growth has been accelerating, is near a historical high point, and is likely physically impossible to continue at this rate much longer. https://www.cold-takes.com/this-cant-go-on/

Moreover, we may witness explosive, unfathomable growth in intelligence and capability in the very near term. The nature of reality may be very different in 10-50 years than it is today. Outcomes ranging from utopia to total human elimination are realistically possible within that timespan. While those extremes are speculative, we will likely see AI-related inventions on par with the modern computer that profoundly impact our lives.

Many technologies are being developed, like AR/VR/MR, cryptocurrency, nanotechnology, quantum computing, various biology-related improvements, fusion power (and other energy solutions), and more. Whether the development speeds of these technologies move at a traditional pace or a breakneck exponential pace will largely depend on the power of AI.

This article is focused primarily on synthesizing the likely timelines and risks of AI, each of which have been hotly debated since the launch of ChatGPT. The smartest and most informed experts are on all sides of the debate. Many express extreme levels of confidence, despite other highly intelligent experts fully understanding their arguments and believing the exact opposite with similar certainty. My personal stance is that there is definitely a non-zero and non-100% probability of many outcomes, based on what we currently know. While someone will turn out to be almost exactly right, perfect confidence is misplaced. I generally believe most of those opining on this topic are sincere and honest in their opinions, but generally have innate biases that are influenced by their professional positions and their prior public statements.

To avoid making this a book, I will provide an overview of the most insightful arguments in a hierarchical structure, with the ability to double-click into any point.

Writing this article served the primary purpose of allowing me to organize my thoughts on this topic so that I could form my own opinion. But it also can be used to educate others or to let others fill in key holes. At the end are specific action items we should aim to take to maximize our expected utility.

One more foundational point: If you haven't been paying attention, this sounds like it's posted by someone who is entirely out of touch with reality -- someone far too invested in science fiction. Some of the potential advancements don't *feel* remotely feasible in the next decade. That's part of

the point: High exponential rates of impact just don't *feel* possible to a layperson and can sneak up fast. If you feel this way, it's important that you become aware of how impactful this technology may be in the relatively near-term.

I. Definitions

Definitions are important only so that we're on the same page. If you want to define AGI as less intelligent than my definition, then the timelines for AGI and the risks from AGI would be reduced. And vice versa.

I. Narrow AI - AI that is skilled in only a specific task or narrow group of tasks.
II. AGI -
    i. Weak AGI: A single AI that is at least equivalent to an average human at nearly all intelligence- and knowledge-related tasks (in a chat only).
    ii. Strong AGI: A single AI that is equivalent to a world class human in practically all intelligence- and knowledge-related tasks/domains (including in the physical world) and is able to play a substantial role in discoveries in physics.
III. ASI - A single AI that is better than the best humans at most or all intelligence- and knowledge-related task.

II. Realistic Powers of AGI

- i. Potential Abilities
    - i. Economic Impact and Labor Market Transformation
        1. Automation of cognitive tasks across industries, potentially leading to significant GDP growth.
        2. Examples:
            a. Financial sector: AGI systems handling complex market analysis, risk assessment, and portfolio management.
            b. Legal industry: AGI conducting legal research, contract analysis, and even basic case preparation.
            c. c) Customer service: AGI providing human-like support across all channels, 24/7.
        3. Underlying abilities: Rapid information processing, natural language understanding, decision-making in complex environments.
    - ii. Scientific Research and Development
        1. Acceleration of scientific discoveries and technological innovations.
        2. Potential for breakthroughs in fields like medicine, materials science, and clean energy.
        3. Example: AGI systems designing and running experiments, analyzing results, and formulating new hypotheses at superhuman speeds.
        4. Underlying abilities: Advanced pattern recognition, hypothesis generation, multidomain knowledge integration.

iii. Healthcare and Medicine
    1. Improved diagnostic accuracy and personalized treatment plans.
    2. Acceleration of drug discovery and development processes.
    3. Example: AGI analyzing patient data, genetic information, and latest research to provide tailored health recommendations.
    4. Underlying abilities: Complex data analysis, pattern recognition in large datasets, causal reasoning.

iv. Education and Skill Development
    1. Personalized learning experiences adapted to individual students' needs and learning styles.
    2. Rapid curriculum updates to keep pace with changing job market demands.
    3. Example: AGI tutors providing one-on-one instruction in any subject, adjusting in real-time to student progress.
    4. Underlying abilities: Natural language processing, adaptive learning, knowledge synthesis across domains.

v. Entertainment & Consumer
    1. Generation of personalized content across various media.
    2. New forms of interactive and immersive entertainment.
    3. Robots performing nearly all household tasks.
    4. Example: AGI creating customized storylines in games or generating music tailored to individual preferences.
    5. Underlying abilities: Creativity, understanding of human emotions and preferences, adaptive content generation.

vi. Environmental Management and Climate Change Mitigation
    1. Development of innovative solutions for renewable energy and carbon capture.
    2. Example: AGI designing highly efficient solar cells or developing new methods for atmospheric carbon extraction.
    3. Underlying abilities: Advanced simulation capabilities, creative problem-solving, integration of multidisciplinary knowledge.

vii. Urban Planning and Infrastructure Management
    1. Optimization of city systems for efficiency, sustainability, and livability.
    2. Predictive maintenance and intelligent resource allocation.
    3. Example: AGI systems managing traffic flow, energy grids, and waste management in real-time; ubiquitous fully autonomous cars.
    4. Underlying abilities: Complex systems modeling, predictive analytics, multi-objective optimization.

viii. Governance and Policy Making
    1. Data-driven policy development and impact assessment.
    2. Improved long-term strategic planning capabilities.

3. Example: AGI systems analyzing vast amounts of socioeconomic data to predict policy outcomes and suggest optimal strategies.
4. Underlying abilities: Complex scenario modeling, causal reasoning, ethical decision-making.

ix. Agriculture and Food Production
1. Precision farming and optimized crop management.
2. Development of new, more resilient and nutritious crop varieties.
3. Example: AGI systems managing entire farms, from soil analysis to harvest timing, maximizing yield while minimizing environmental impact.
4. Underlying abilities: Multivariable optimization, predictive modeling, integration of biological and environmental data.

x. Space Exploration and Resource Utilization
1. Advanced mission planning and execution.
2. Autonomous space vehicle and habitat management.
3. Example: AGI systems designing and controlling self-replicating machines for extraterrestrial resource extraction.
4. Underlying abilities: Long-term strategic planning, autonomous decision-making in unknown environments, advanced physics and engineering knowledge.

xi. Sources
1. https://www.forbes.com/sites/cindygordon/2023/09/30/how-general-ai-will-eventually-reshape-everything/
2. https://www.aei.org/articles/ai-and-the-economy-scenarios-for-a-world-with-artificial-general-intelligence/
3. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi
4. https://www.linkedin.com/pulse/unlocking-future-artificial-general-intelligence-its-implications-t/

ii. Impact Potential of Generative AI
i. Obviously, if generative AI *does* lead to AGI, the impact potential is huge. And this may very well be the path.
ii. If not, generative AI is still *useful*, even at its current quality. If we see improvements in GPT-5 and GPT-6, the utility will increase substantially.
iii. ChatGPT itself is helpful. But that has fairly low limits on economic impact and utility.
iv. Instead, purpose-built application level software will be very useful over the next couple of years - Sam Altman has come to this realization himself (and this is probably part of why he's making ChatGPT with GPT4o free - he is going to make money off the API calls, which are good margin, whereas the cost to run ChatGPT is ~$2/user/year).
v. As a side note, some of the keys for application developers are:

1. Use reliable external data and retrieval augmented generation (RAG) techniques, along with anti-hallucination loops, to limit/eliminate hallucinations.
2. If the AI is imperfect, craft flows to keep the user in the loop in a visual way to make decisions, and let the AI handle take on tasks it performs best (like processing of huge amounts of data and finding needles in a haystack).
3. Craft powerful agentic flows (with each endpoint using the ideal foundation model) that provide good experiences by understanding the needs of the user and the current capabilities of technology - for example, you may have a 10-step flow but start working on a very complicated Step 9 in the background right after Step 3 begins.
4. Show work/sources to the user for verification.
5. Focus on frictionlessly bringing in necessary context for the AI to make decisions.
6. Think through the full end-to-end workflow to create real value.

vi. But, developing for LLMs is certainly challenging. Generative AI engineer requires a higher-skilled engineer than traditional programming. The error checking is *much* harder. API rate limits and timeouts still occur frequently. And every now and then, a generally good result goes way off course.

vii. Deep tech - the layer between the foundation models and application-level software (like RAG architectures/tools and similar) is also improving rapidly. This allows developers to generate significantly more value from the current foundation models.

iii. Synthesis

i. These implications suggest that AGI could drive enormous productivity gains across multiple sectors, potentially leading to a new era of economic growth and scientific advancement.

ii. However, the implications also highlight the need for careful consideration of the societal impacts, including potential job displacement and the need for reskilling programs.

III. Realistic Powers of ASI

i. Singularity-level superintelligence is certainly not guaranteed. And, a superintelligent agent might provide negative utility to humans. It's helpful, nonetheless, to understand the potential power of a realistically possible ASI.

ii. As detailed by Nick Bostrom, ASI can have immense powers at technological maturity, very likely *at least* including the following abilities:

i. Cures for all diseases / Reversal of aging

ii. High-throughput atomically precise manufacturing*

iii. Realistic simulations of reality

iv. Von Neumann Probes (self-replicating space colonization machines that can travel at a substantial fraction of the speed of light) / Dyson spheres (for harvesting the energy output of stars)
v. Precision control of the mind (e.g., having the positives of the most intense psychedelic trip experienced)
vi. And much more:
    1. Manufacturing & robotics
        a. High-throughput atomically precise manufacturing
        b. Distributed robotics systems at various scales, including with molecular-scale actuators
    2. Artificial intelligence
        a. Machine superintelligence that vastly exceeds human abilities in all cognitive domains
        b. Precision-engineered AI motivation
    3. Transportation & aerospace
        a. von Neumann Probes (self-replicating space colonization machines that can travel at a substantial fraction of the speed of light)
        b. Space habitats (e.g., terraforming suitable planets or free-floating platforms such as O'Neill cylinders)
        c. Dyson spheres (for harvesting the energy output of stars)
    4. Virtual reality & computation
        a. Realistic simulations (of realities that to human-level occupants are indistinguishable from physical reality, or of rich multimodal alternative fantasy worlds)
        b. Arbitrary sensory inputs
        c. Computer hardware of sufficient efficiency to enable terrestrial resources to implement vast numbers of fast superintelligences and ancestor simulations
    5. Medicine & biology
        a. Cures for all diseases
        b. Reversal of aging
        c. Reanimation of cryonics patients
        d. Full control of genetics and reproduction
        e. Redesign of organisms and ecosystems
    6. Mind engineering
        a. Cognitive enhancement
        b. Precision-control of hedonic states, motivation, mood, personality, focus, etc.
        c. High-bandwidth brain-computer interconnects
        d. Many forms of biological brain editing
        e. Digital minds that are conscious, in many varieties
        f. Uploading of biological brains into computers
    7. Sensors & security
        a. Ubiquitous fine-grained real-time multi-sensor monitoring and interpretation

b. Error-free replication of critical robotic and AI control systems
c. Aligned police-bots and automatic treaty enforcement

8. More speculatively is the question of whether ASI would be able to find end-runs around what are currently believed to be constraints of physics, like the universal speed limit, the ability to stop the heat death of the universe, the ability to create/use wormholes, and similar.
9. See: Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014. Available on Amazon. https://www.amazon.com/Superintelligence-Paths-Dangers-Strategies-Nick-Bostrom/dp/0199678111

iii. On the other hand, there are serious questions regarding whether ASI might be unable to substantially accelerate technological progress or make breakthroughs in fundamental research. This depends on: (1) how much of its potential intelligence the ASI has reached; and (2) how grand the potential is for intelligence.
- For example, it may be that (professor-level) human thought - while perhaps not maximally fast - is about as sophisticated as reasoning gets.
- And, even if reasoning can get more sophisticated, it may be impractical to just intuit the laws of physics (for example) - we may still need to run the experiments.
- On the other hand, if we had 10,000 Einsteins, all equipped with all human knowledge and able to process at 10,000X speed, there may be some amazing breakthroughs very quickly - even if it does still require running experiments (which can be very surgical, with exact instructions given on how to set them up quickly).

- **Another major question regards AI consciousness - subjective qualia.**
  - **We do not yet have a good idea of what causes consciousness. Most philosophers and scientists are materialists/physicalists (rather than dualists) and believe consciousness arises from a certain (unknown) pattern of brain activity. Perhaps recursive feedback loops or quantum entanglement are key. Perhaps the overall power of the brain becomes relevant. Given that the brain is effectively just a computer, machines will, in all likelihood, be capable of becoming conscious, too.**
  - **There is a wide spectrum of subjective conscious experience. The experience of an insect or a bat is a fraction as rich as the experience of a human. Extremely advanced machines will likely have subjective conscious experiences (qualia) that are vastly richer than the experiences of humans — perhaps with many more senses, a deeper sense of meaning, greater levels of feeling, and more. But, given the difficulty in measuring consciousness, we *might* never be sure consciousness has arisen in our machines, since even the LLMs of today can fake**

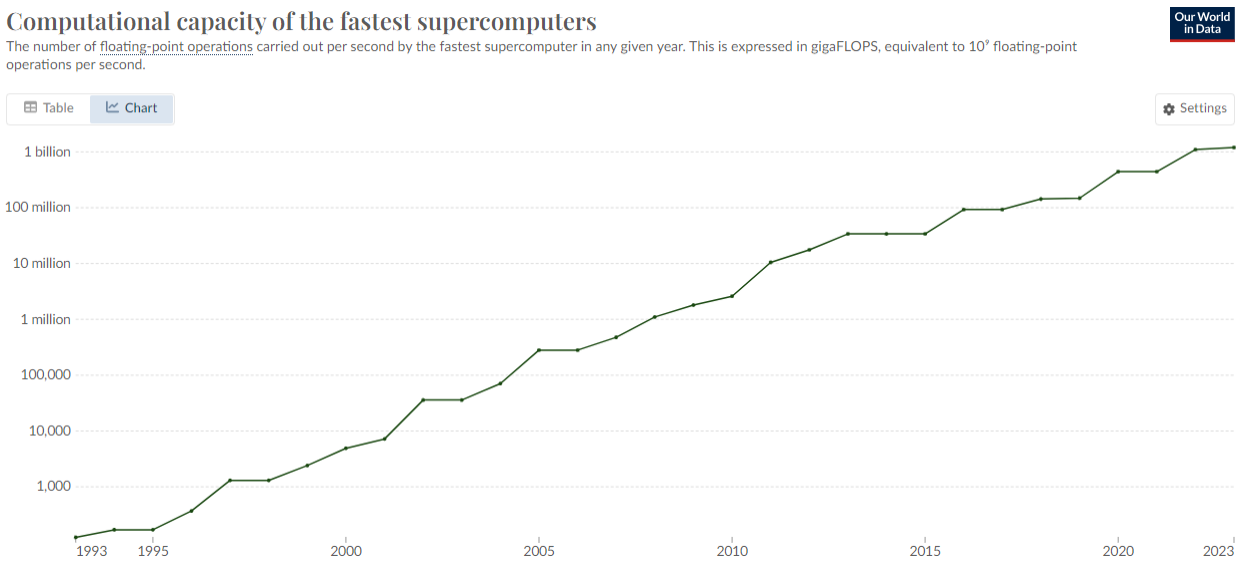
consciousness. I suspect, though, there will be a few ways to figure this out.

IV. Speed of AI Advancement

Onto the heart of this article: When will we have weak/strong ASI? Let's analyze the facts.

I. Background / Initial Progress
   - **i.** Began meaningfully discussing the concept of AI in the 1950s - at times thinking we were close to developing it.
   - ii. In the past, the assumption was that we'd reach AGI via a traditional computer program. That *is* still possible (and a weak ASI may be able to develop such a program).
   - iii. But current state-of-the-art (SOTA) methods involve extreme matrix multiplication by weighting training sets of enormous repositories of data.
   - iv. Processing power is an important factor.
     - i. Open question whether machines will need to have human-level processing to reach human levels of intelligence.
     - ii. Also an open question whether they will need to be optimized for the same level of efficiency as the human brain.

V.



VI. Once a computer is able to solve nearly all the problems that humans can, it will likely be vastly more intellectually capable than humans.
   - i. A simple calculator is much better than humans at arithmetic.
   - ii. Computers are already better than humans at the most complex games (chess, Go, etc.).
   - iii. Computers can analyze large amounts of numerical data and find trends much faster than humans.
   - iv. Computers can store large amounts of abstract information in memory (humans can, too, but it involves a different type of data).

VII. The consensus estimate through around 2014 was that the fastest path toward AI would involve cloning the human brain digitally (there is no scientific reason to think we can't reproduce a brain in silicon).
    i. Brain imaging is moving slower than we hoped, however, and we now see much faster traction using deep learning methods.
    ii. Full brain emulation is not expected until around 2075, and that may be partially aided by advances in AI. Prediction markets suggest that there is a 98% chance that an approach that is different from brain emulation will yield the first AGI.
        i. https://www.metaculus.com/questions/2813/date-of-first-human-whole-brain-emulation/
        ii. https://www.metaculus.com/questions/372/will-human-brain-emulation-be-the-first-successful-route-to-human-level-digital-intelligence/
    iii. As a side note on this point: Unless the dualists end up being right (and maybe even if they are), given that the brain is a machine that can be recreated in silicon, we'll almost certainly be able to create machine consciousness eventually.
    iv. Those who say ASI will *never* be a threat because it will always just be a tool or computer program are probably unfairly dismissing this point.

VIII. https://en.wikipedia.org/wiki/Artificial_intelligence

II. Recent Progress
    i. Up until GPT-3, the consensus estimate on the arrival year of AGI was 2045 (and a decade earlier, it was 2065). Shortly after ChatGPT, the estimate was 2032. Forecasters were very surprised by recent AI progress.
    ii. Built on deep learning techniques, particularly transformer architectures, large language models (LLMs) have emerged as a revolutionary approach in AI. These models employ matrix multiplication and a weighting of (vast) training datasets to predict the next token (basically, the next word) in a sequence. While this description may seem reductive, it belies the profound complexity and elegance of the technique. By scaling up this approach, we've witnessed the emergence of capabilities that appear to transcend the model's basic training objective, resulting in what many consider a breakthrough in artificial intelligence. This emergent intelligence, arising from a conceptually straightforward yet computationally intensive method, has reshaped our understanding of what's possible in machine learning and natural language processing.
        i. **https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf**
    iii. The question is: How much can we scale this technique? This comes down to our collective ability to improve across 3 domains:
        i. Scaling compute power via extreme training runs (GPT-5 is likely in the ~$1-3B range, and it's possible we will scale to $100B or $1T for future training runs).
            1. This is very expensive from an energy perspective and raises some climate change concerns.
            2. On the other hand, AI may ultimately help us solve climate change.

3. Aside from energy, this also depends on the availability and power of GPUs, and NVIDIA in particular has seen tremendous growth based on its technical superiority in an area with extreme demand.
4. This is slowed, in the short-term, by a bottleneck with NVIDIA that is, in particular, slowing down OpenAI relative to xAI and Meta.
    a. **https://www.wired.com/story/nvidia-chip-shortages-leave-ai-startups-scrambling-for-computing-power/**

ii. Increasing the availability of training data
1. One element includes acquiring wide amounts of high-quality human-generated text.
    a. This has become legally contentious, since the top research groups likely used extensive copywritten underlying data to train their models without authorization. But, they were often using data they paid to access.
    b. Now, groups are protesting this usage (which is opaque from a legal perspective) and many sites are preventing bots from crawling the site through both their terms of service and by erecting technical barriers.
    c. Given the race to AI, ignoring some of the more burdensome rules could provide a tactical advantage in speed.
2. As AI becomes better, there will likely be increasing amounts of AI-generated "synthetic" training data.

iii. Improving the algorithms
1. There are many creative tricks to continuously improve the sophistication of the foundation models.
2. For example, instead of simply predicting the "next token," a research paper from Meta suggested predicting the next group of tokens.
    a. **https://arxiv.org/pdf/2404.19737**
3. And OpenAI had a major breakthrough last year, which is believed to be linked to Q* reasoning (project Strawberry).
    a. **https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12/**
4. Another major improvement comes from Mixture-of-Experts (MoE) architectures, which is used by Google Gemini Pro 1.5.
    a. **https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/**
5. Many engineers have also found potential paths to make the AI less of a "black box" and use that to reduce hallucinations.

6. Some also expect the LLMs to have a "fast answer" and "slow answer" method of approaching problems, with harder problems requiring more analysis.
7. Researchers have also found ways to use larger, intelligent models to create very impactful, small training sets for smaller models so that much cheaper/faster models are not too far from the leading foundation models.
8. These seem like the biggest near-term breakthroughs, and they're likely very impactful.

iv. https://lexfridman.com/sam-altman-2-transcript

iv. Will these techniques be highly scalable?
   i. The hotly-debated consensus answer is *probably* yes, to a point, but scaling existing methods (with some algorithm improvements) might not take us to AGI.
   ii. It's certainly possible that there is not much difference in intelligence between a 900B parameter model and a 40T parameter model - we may have already maxed out the useful size of the training runs.
   iii. But, regardless, there is likely a limit to the realistic expense that can be spared for a training run. Maybe it is $100B, and maybe it is $1T if a breakthrough is very likely at that price point. But, much beyond that level is not realistic in the near/mid-term.

v. There are also major questions about whether large language models - or auto-regressive models altogether - are the right path to superintelligence.
   i. The architecture may fundamentally be unable to remove hallucinations.
      1. Side note: While this is true, RAG techniques can be used to almost fully eliminate hallucinations in application-level software, but properly crafting the experience is key for usability/speed.
   ii. Critics note that LLMs are simply "stochastic parrots" without actual thinking/reasoning capabilities.
      1. For example, on certain math and reasoning capabilities, SOTA LLMs still fail spectacularly. A common example can be shown by an LLM's response to the following prompt, "A farmer and a sheep are on one side of a river, with one raft. The raft can hold the farmer and only one other animal while helping to help cross the river. How can the farmer cross the river in as few trips as possible while arriving safely on the other side?"
      2. Simple arithmetic (multiplying 4-digit numbers) and more advanced math challenges prove difficult for current LLMs.
         a. https://www.linkedin.com/pulse/911-greater-than-99-conversation-genai-abhinav-saxena-ycirf/
      3. OpenAI is attempting to fix this via 'Q* Reasoning,' which they've codenamed Strawberry, but the fix is not guaranteed to work.

4. My own take is that we shouldn't worry too much about terms like 'thinking' or 'reasoning,' nor assume an AI is 'smart' or 'dumb.'
    a. Before AI is conscious, it shouldn't be thought of as an entity like humans.
    b. Instead, we should focus on the empirical results - what types of problems can the AI solve consistently, inconsistently, and not at all?
    c. "Cheap tricks" might create emergent behavior and might very well be the path to creating highly useful tools *or* even AGI.
    d. But, clearly, there are still some gaps to fill, and it's not obvious whether these are fillable via LLM architecture.
5. **https://www.lesswrong.com/posts/HxRjHq3QG8vcYy4yy/the-stochastic-parrot-hypothesis-is-debatable-for-the-last**
6. **https://pli.princeton.edu/blog/2023/are-language-models-mere-stochastic-parrots-skillmix-test-says-no**

vi. OpenAI representatives have said that SORA, their video-based generative model, is the likely path to AGI.
    i. Some of the latest technologies previewed or released from the top research labs have shown video, vision, hearing, sound, and omni models that are advancing rapidly in abilities. Given the amount of data contained in a video vs. text, this may indeed be the key.
    ii. "Sora serves as a foundation for models that can understand and simulate the real world, a capability we believe will be an important milestone for achieving AGI."
        1. **https://openai.com/index/sora/**

III. Progress on frontier intelligence models has been slow since the release of GPT-4.
    i. The argument is that, since February 2024, we have had only very modest increases in the intelligence of the leading SOTA model (which, as of today, is either Anthropic's Sonnet 3.5 or OpenAI's GPT4o).
    ii. While it is true that the research labs have been able to retain GPT-4-level intelligence at a lower model size, which means lower cost and better speed, they have not shown much more than trivial increases in intelligence.
    iii. And, many of these increases in intelligence are based on benchmarks, but there is some gaming involved - many developers have questioned whether any model has surpassed the original GPT-4 (although Sonnet appears to be well-respected for it's coding skills and seems to be accepted as an improvement).
    iv. But, this perceived slow rate of progress is an unfair characterization. The release cycle from OpenAI has been every 34 months for a +1 model release. We are still within that window. Although they have much more resources than in the past, the difficulty of improvement is much higher than it was in the past, so more resources are expected.
    v. While it does appear that the rate of progress isn't going to be moving at breakneck speed the next 6 months, it isn't clear that speed has slowed.

vi. **https://artificialanalysis.ai/models**

IV. How does the current level and trend of improvement in processing power align to the development of AGI?
   i. Estimating the processing power of the human brain is difficult and the confidence ranges are wide. Converting those estimates to FLOPS (a supercomputer processing metric - floating point operations per second) yields an estimate of human brain processing power being 10^{12} to 10^{28} FLOPS (most likely 10^{18} to 10^{25}). Remember that these are huge differences - 10^{25} is 10,000,000X more powerful than 10^{18} (I'm surprised we can't predict more accurately).
   ii. Given the vast amount of resources being poured into AI-related supercomputers, the top system in 2025 is expected to have $3x10^{26}$ FLOPs. That should put it on par with at least 300 human brains.
      i. https://www.metaculus.com/questions/12937/greatest-computation-used-in-ai-training/
   iii. The human brain is *highly* efficient, however, and current LLM architectures use more of a brute force approach for the largest models (and then often distill intelligence in a much more efficient way into the smaller models), so we may still have work to do.
      i. Watson used ~85,000 watts to win Jeopardy. ChatGPT is probably 1-2 orders of magnitude less efficient than the human brain, *after* a highly energy intensive run.
      ii. https://www.humanbrainproject.eu/en/follow-hbp/news/2023/09/04/learning-brain-make-ai-more-energy-efficient/
      iii. https://arxiv.org/abs/1602.04019
      iv. https://ai.stackexchange.com/questions/38970/how-much-energy-consumption-is-involved-in-chat-gpt-responses-being-generated

V. How far along is OpenAI?
   i. The answer here is that very few people outside of OpenAI know how far along OpenAI really is, and estimates vary wildly. This is important, because if OpenAI is stalling out already (which is possible), the timeline for AGI changes drastically.
   ii. OpenAI has done a spectacular job of being secretive, and many rumors have proven to be completely wrong. Let's take a look at the facts we do know that might give some insight.
   iii. Business Insider released an article suggesting that OpenAI was likely set to release GPT-5, which was a huge advancement over GPT-4, over the summer of 2024, citing anonymous CEOs who tested the new model. Note, however, that Business Insider likely lacks subject matter expertise and has seen several very public flags raised about its credibility and truthfulness.
      i. https://www.businessinsider.com/openai-launch-better-gpt-5-chatbot-2024-3
   iv. Mira Murati, the CTO of OpenAI, has made a few public comments in recent weeks:
      i. Next foundation model will advance from a smart high-schooler (GPT-4) to a PhD student in intelligence and will be released in 1.5-2 years (mid-2024 statement). But, other sources suggest that in April

2024, OpenAI told internal employees that it was "on the cusp" of models capable of "problem-solving tasks as well as a human with a doctorate-level education."
    1. https://x.com/tsarnick/status/1803901130130497952
    2. https://x.com/AISafetyMemes/status/1811579385222475960
  ii. OpenAI has released to the public nearly all of its internal capabilities.
    1. https://x.com/tsarnick/status/1801022339162800336
v. Sam Altman interviews:
  i. On Lex Fridman in March 2024:
    1. "I expect that the delta between 5 and 4 will be the same as between 4 and 3."
    2. "We will release an amazing new model this year. I don't know what we'll call it."
    3. https://lexfridman.com/sam-altman-2-transcript
  ii. GPT-4 is "The dumbest model any of you will ever have to use, by a lot."
    1. https://www.youtube.com/watch?v=fFIiIp8ZrDg
  iii. "We can say right now, with a high degree of scientific certainty, GPT-5 is going to be a lot smarter than GPT-4 and GPT-6 will be a lot smarter than GPT-5, we are not near the top of this curve."
    1. https://x.com/ns123abc/status/1783360235010199867
vi. OpenAI publicly announced that it began training its next frontier model around May 2024 (although it has not been 100% clear whether this is GPT-5 in training vs. GPT-6 - it could be that GPT-5 is further along).
  i. https://openai.com/index/openai-board-forms-safety-and-security-committee/
vii. Historically, OpenAI released a new foundation model every 34 months, which aligns well to Mira's estimate, with GPT-4 released in February 2023 and GPT-5 being released around December 2025.
viii. Outside of direct statements, we can gain some insight based on actions:
  i. An OpenAI engineer reportedly sold his San Francisco house because he believes ASI is imminent (presumably, he thinks ASI will lead to a singularity).
    1. https://x.com/jordnb/status/1807938650027745583
  ii. Ilya left OpenAI to start his own ASI company - *not* an AGI company - and doesn't plan to release any product until ASI is reached. He expects to have at least $1B in funding.
    1. https://fortune.com/2024/06/20/openai-ilya-sutskever-sam-altman-safe-superintelligence/
ix. OpenAI has been working closely with the Department of Defense (DoD). This could have some implications on the future of warfare, but also, some speculate that the DoD is receiving the cutting edge breakthroughs and slowing OpenAI's release cycle. OpenAI's terms prohibit using the tech for weapons, however, and OpenAI is publicly stating that it is helping improve cybersecurity capabilities.

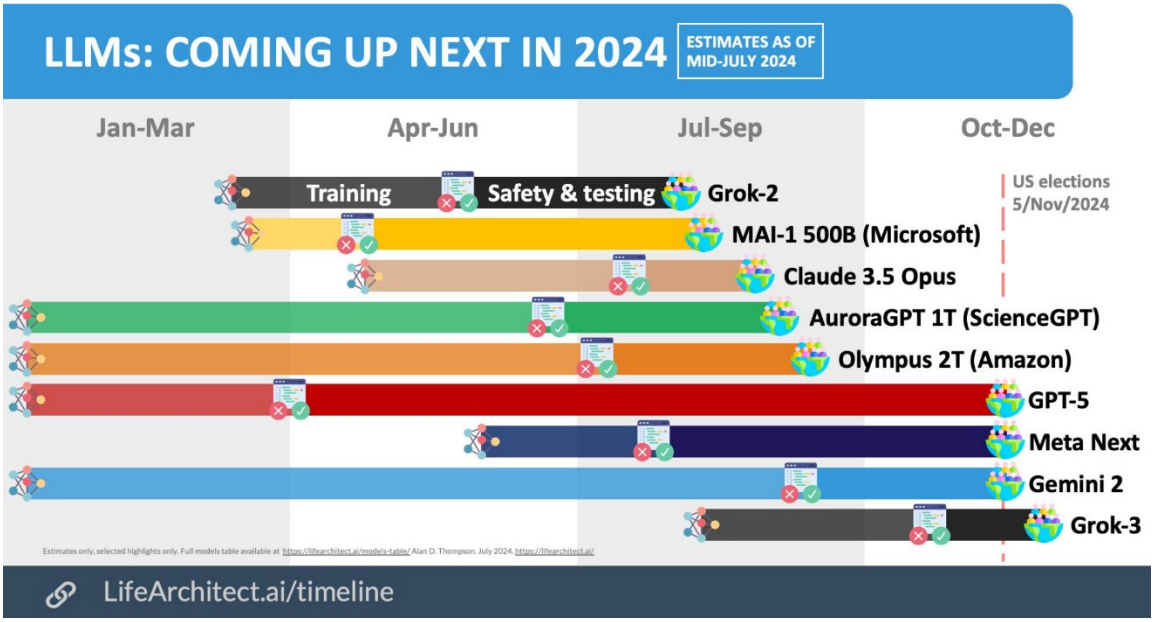
i. https://www.bloomberg.com/news/articles/2024-01-16/openai-working-with-us-military-on-cybersecurity-tools-for-veterans?embedded-checkout=true

x. Some commentators question Sam Altman's candor. The argument is that, given that he has been fundraising huge amounts of money, it's important for him to show inevitable and (ideally) near-term progress. Comparatively, Mark Zuckerberg has much longer timelines (although my guess is Mark is heavily influenced by Yann LeCun who just has a unique perspective).

xi. OpenAI also has been struggling to release the voice mode for its Omni model, which has made many question whether it can still ship. But, this is likely driven by legal/copyright concerns.

xii. Conducting an enormous $1B+ (including equipment) training run is extremely difficult logistically, specifically including acquiring sufficient numbers of GPUs. It appears this is the main bottleneck.

xiii. A well-documented collection of rumors from an anonymous industry expert suggests that Q4 2025 could be a time when many companies release true next-gen models. It could also prove to be entirely too optimistic and factually wrong.



i. https://x.com/koltregaskes/status/1818419846755107053

ii. Recent update: Some of this may be delayed based on the recent NVIDIA shortage.

VI. Modest continuous intelligence increases are exponential in impact.

i. A high school level intelligence that hallucinates, especially if it is slow and does not have proper workflows set up to gain context, probably can't provide too much economic value, because the end users are smarter than the AI.

ii. But, once we see PhD level intelligence *and* application-level companies have built sophisticated workflows to bring in key context, nearly every job becomes substantially automated.

iii. If LLMs can solve hallucination issues, they become *substantially* more valuable than they are today. But 100% is much better than 99.5%.

i. Prediction markets indicate that by June 2025, we may see 1/5 as many hallucinations as we did when GPT-4 was first released. **https://www.metaculus.com/questions/17443/5x-less-gpt-hallucination-by-june-30-2025/**

ii. Some researchers think reducing (or perhaps eliminating) hallucinations is possible. **https://x.com/DanHendrycks/status/1709227490592612671?t=NEuLnI039GHaNq6z1WihNw&s=19**

VII. Prediction Markets

i. Consensus GPT-5 release date: March 2025 (announcement date of January 2025)

i. https://www.metaculus.com/questions/15462/gpt-5-announcement/

ii. https://www.metaculus.com/questions/22047/when-will-gpt-5-be-publicly-available/

iii. But, if legislation limiting LLMs is introduced, GPT-5 may be delayed until early 2026.

1. https://www.metaculus.com/questions/17170/conditional-gpt-5-announcement/

iv. GPT-5 vs GPT-4

ii. Likelihood OpenAI will claim GPT-5 is AGI: 2%.

i. https://www.metaculus.com/questions/18360/gpt-5-agi-or-not/

iii. By July 2025, the top LLMs will hallucinate 1/5 as much as GPT-4 did when released.

i. https://www.metaculus.com/questions/17443/5x-less-gpt-hallucination-by-june-30-2025/

iv. 8% (OpenAI GPT-5) to 20% (Google Gemini Ultra 2) likelihood that AI capabilities will plateau after the release of the company's *next* frontier model.

i. **https://www.metaculus.com/questions/22519/ai-capabilities-plateau-2025/**

VIII. Is Generative AI the answer for AGI?

i. Some of the leading research scientists, like Yann LeCun of Meta, believe AI is at least a decade away and that LLMs are not the right path.

i. Yann does not think LLMs or auto-regressive models are the proper path to AGI. Prediction-based language models will inherently struggle to fundamentally reason.

ii. In addition to inherent issues with hallucinations, LLMs lack a true 'world model,' which Yann and others believe is a requirement for true reasoning capabilities.

iii. His path to AGI, which is probably not entirely baked, is something like the following:

1. Start with vision, rather than language - vision requires orders of magnitude more data.

2. Continue to scale up processing power - the human brain is *much* more efficient and still orders of magnitude higher in power.
3. Work on increasingly improving levels of abstraction - we should be able to see a car and understand that it moves.
4. Move away from generative/auto-regressive models and toward self-supervised learning and energy-based planning models.
    a. Gradient descent can help find ideal shapes of output.
    b. The AI should decide whether questions are hard or easy and assign the appropriate amount of energy accordingly.
    c. Create predictive world models that can build internal models of the world and make predictions about future states, rather than just pattern matching on past data.
    d. Pursue planning models that evaluate questions and form hierarchical plans for solutions.
    e. Develop joint embedding architectures that can learn to represent different types of data (text, images, etc.) in a shared latent space, allowing for more flexible and powerful reasoning across modalities.
    f. **https://lexfridman.com/yann-lecun-3-transcript**

iv. There are other frameworks supported by other AI experts:
1. Hybrid models AlphaProof & AlphaGeometry2 achieved a silver-medal standard on Math Olympiad problems, which are great examples of formal reasoning problems where LLMs struggle.
    a. The systems translate problems into the formal mathematics language Lean.
    b. AlphaProof uses reinforcement learning algorithms (like those used by AlphaZero).
    c. By using the formal mathematical language, the proofs can be formally verified for correctness.
    d. When presented with a problem, AlphaProof generates solution candidates (a generative approach) and then proves or disproves them by searching over possible proof steps in Lean. Each proof that was found and verified is used to reinforce AlphaProof's language model, enhancing its ability to solve subsequent, more challenging problems.
    e. The generated solution candidates are created using a neural network model.
    f. At a high-level, this is a clever hybrid model, mixing generative steps, reinforcement learning, and formal logic.

    g. **https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/#:~:text=When%20presented%20with%20a%20problem,solve%20subsequent%2C%20more%20challenging%20problems**.

2. Neurosymbolic reasoning, more generally, may prove to be a useful approach that solves hallucination issues.
    a. Neurosymbolic approaches combine neural networks (which excel at pattern recognition and learning from data) with symbolic reasoning (which handles logic, rules, and structured knowledge).
    b. This hybrid method leverages the strengths of both approaches: neural networks for perception and data-driven tasks, and symbolic reasoning for logical inference and understanding structured knowledge.
    c. By integrating these methods, neurosymbolic systems can address complex problems that require both learning from examples and applying formal reasoning, such as understanding natural language or solving mathematical proofs.
3. As we learn more, we may realize that AGI is harder to achieve than we expected. A difficult but likely successful approach would involve implementing architectures that resemble the framework of the human brain.
    a. Deep learning, at a very high level, does this already. But it does not go nearly far enough. We may need to more closely mimic the human brain's architectural features.
    b. Feedback Loops: Mimic the brain's ability to learn from past experiences and adjust responses in real-time, enhancing decision-making processes and adaptability.
    c. Neuron-like Architectures: Develop artificial neurons and synapses that replicate the brain's structure, enabling more efficient and complex information processing.
    d. Parallel Processing: Implement parallel processing capabilities to allow simultaneous handling of multiple tasks, similar to how the brain processes different types of sensory input concurrently.
    e. Plasticity: Incorporate mechanisms for synaptic plasticity, allowing the system to rewire and adapt its connections based on new information and experiences, enhancing learning and memory.
    f. Embodiment: Integrate sensory and motor systems to provide a physical context for the AI, improving its interaction with and understanding of the real world through embodied cognition.

g. Hierarchical Organization: Create layered networks that process information in stages, resembling the hierarchical organization of the brain, to manage complexity and extract higher-level abstractions.

ii. In the case that LLMs fail to produce AGI, alternative approaches showcasing a better path may present more ideal solutions. These alternative architectures will likely be built somewhat in parallel with generative models, but if generative models stall out, significant resources can be redeployed to approaches showing traction.

IX. Betting Markets & Expert Predictions (**make a timeline of events based on consensus predictions**)

i. Metaculus:

i. Weak AGI - 2028 (up from 2026)
   1. Neither strong nor weak AGI exactly match my definition above.
   2. https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/

ii. Strong(er) AGI - 2033
   1. https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/

iii. ASI
   1. It is gone now, but a prior prediction market (in 2023) suggested that ASI would be ~2036, which at the time was ~4 years after AGI.

iv. Weak AGI-to-ASI Oracle timeline: 21 months
   1. Note: This seems inconsistent with some of the other estimates.
   2. https://www.metaculus.com/questions/4123/time-between-weak-agi-and-oracle-asi/

v. Note that, of course, heavy regulation, world wars, and similar would change timelines.

vi. Estimated FLOPs in largest AI training run in 2025: $3x10^{26}$ (and 200X that by 2032).
   1. https://www.metaculus.com/questions/12937/greatest-computation-used-in-ai-training/

vii. Reliable & general household robots *developed* by 2035 (but not for sale to general public for ~8 more years), and humanoid robots indistinguishable from humans by 2055.
   1. https://www.metaculus.com/questions/16625/date-of-reliable-and-general-household-robots/
   2. https://www.metaculus.com/questions/24832/diaper-changing-robot-availability-date/
   3. https://www.metaculus.com/questions/15465/robots-identical-to-humans/

viii. Cost of most expensive training run, by year:
   1. 2024: $433M
   2. 2026: $771M
   3. 2030: $1B

4. Note: This is just for a single training run. The infrastructure costs are many multiples more expensive. ~$544B is expected to be invested in AI companies in 2025.
5. https://www.metaculus.com/questions/17418/most-expensive-ai-training-run-by-year/

ii. Named Experts (**show chart**)

i. Wide range of expert opinions, with most being 2025-2040, and the consensus median around 2032.

1. **https://x.com/AISafetyMemes/status/1743653636448600532**

X. Synthesis

i. Overall, a major question-mark on OAI; probably best to go with the prediction markets - March 2024 for GPT-5, but some uncertainty.

i. Ignoring the prediction markets, I could be very wrong, but my estimate is that OpenAI will release some smaller improvements through mid-2025, based mostly on algorithmic improvements that fix some holes and drive some improvements from GPT-4. I expect GPT-5 to be released around the end of 2025, and that will be a meaningful step similar to GPT-4 vs. GPT-3. This level of intelligence will be *very* powerful, especially for secondary applications.

ii. But likely still room for noticeable improvement in GPT-5 and GPT-6, which will likely not be (strong) AGI but, with application-level software, will be able to automate a large portion of the workforce.

iii. A bit worse than a coin flip whether generative AI can be scaled to reach "strong" AGI.

i. If yes, 2027-2034 is very realistic for AGI.

ii. If not, 2032-2045 or longer is more realistic.

iv. But, with high confidence, eventually AGI will be achieved if we don't kill ourselves (via nuclear war or similar).

V. Non-Existential AI Risk

i. Economic disruption and job displacement

i. Automation of specific roles like data entry, customer service, and basic analysis

ii. Short-term unemployment in sectors like transportation (e.g., self-driving vehicles)

iii. Shift in job market demands, requiring workers to adapt to new, AI-augmented roles

iv. https://www.brookings.edu/articles/automation-and-artificial-intelligence-how-machines-affect-people-and-places/

ii. Ethical and bias issues

i. AI systems making biased lending decisions or unfairly assessing job candidates

ii. Balancing act between avoiding discrimination and preventing overcorrection that imposes ideological biases

iii. Example: An AI recruitment tool preferring male candidates vs. an overcorrected version that artificially boosts diversity at the expense of merit.

    iv. https://www.apa.org/monitor/2024/04/addressing-equity-ethics-artificial-intelligence
iii. Social media bots and misinformation
    i. AI-generated fake news articles that are increasingly difficult to distinguish from genuine reporting
    ii. Automated bot accounts amplifying divisive content to increase engagement
    iii. Challenges for platforms in moderating AI-generated content without over-censorship
    iv. https://medium.com/friction-burns/lets-talk-about-bots-baby-5e0bfea518a4
iv. Autonomous drones and warfare
    i. Potential for "swarm" attacks using numerous small, AI-guided drones
    ii. Risk of accidental escalation due to autonomous systems misinterpreting situations
    iii. Debate over human accountability in AI-driven military decisions
    iv. https://www.bloomberg.com/opinion/articles/2024-03-12/don-t-fear-ai-in-war-fear-autonomous-weapons?embedded-checkout=true
v. Privacy concerns
    i. AI-powered facial recognition in public spaces leading to loss of anonymity.
    ii. Advanced data analysis allowing companies to infer sensitive personal information.
    iii. Tension between data-driven services and individual privacy rights.
vi. Security vulnerabilities
    i. AI-powered password cracking and network penetration tools.
    ii. Sophisticated spear-phishing attacks using AI-generated personalized content.
    iii. Potential for adversarial attacks on AI systems themselves (e.g., fooling self-driving car sensors).
    iv. https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/
vii. Power usage
    i. Increased electricity demand from data centers running AI models.
    ii. Local grid strain in areas with high concentration of AI research facilities.
    iii. Need to balance AI advancement with energy efficiency improvements.
    iv. https://www.govtech.com/blogs/lohrmann-on-cybersecurity/ais-energy-appetite-challenges-for-our-future-electricity-supply
viii. Nefarious users (biological/chemical weapons, sophisticated hacks)
    i. AI potentially being used to optimize drug formulations for both beneficial and harmful purposes.
    ii. Advanced language models assisting in creating more convincing scams or malware.

iii. Challenge of restricting AI capabilities without hindering legitimate scientific research.
iv. https://www.safe.ai/ai-risk

VI. Existential AI Risk

i. Factors Affecting Existential Risk

Lower Risk	Category	Higher Risk
Low	**Actual Potential of ASI**	High
Powerful Narrow AIs	**Breadth of AI**	Broad AGI
Minimal	**Safeguards**	Thorough
Slow Takeoff	**Takeoff Speed**	Fast Takeoff
Oracle	**Type of AGI**	Conscious Entity
Low Competition	**Competition/Races**	High Competition
No	**Access to the Internet**	Yes
Cannot access	**Access to its own "Code"**	Can access & edit
Open Source	**Early Stages: Open/Closed Source**	Closed Source
Closed Source	**Later Stages: Open/Closed Source**	Open Source
High Similarity	**Similarity to Human Intelligence**	Low Similarity
Consolidated	**AI Capability Fragmentation**	Diversified

Blue = >80% likelihood this will reflect reality

II. Analyzing the factors
  i. Actual Potential of ASI
    i. If the ceiling for ASI is low (it's impossible for ASI to become much more powerful than humans), then we have practically *no* risk.
    ii. The more powerful ASI can be, the more room it has to manipulate, control, or discard humans.
    iii. https://en.wikipedia.org/wiki/Technological_singularity
  ii. Breadth of AI
    i. Powerful narrow AIs can have very useful applications (see, for example, AlphaFold) and do not create material existential risk.
    ii. Broad AGI (that leads to broad ASI) is much more dangerous.
    iii. Practically, requiring a wide mix of narrow AI likely has much lower commercial value (and is very difficult to implement in a commercial setting) relative to a more powerful general AI. So there is a trade-off between risk and usefulness.
    iv. https://ubiai.tools/exploring-ai-the-distinction-between-narrow-and-general-ai/
  iii. Safeguards

i. The more safeguards we place on systems, obviously the lower the risk. But, realistically, this also slows development substantially.
ii. Safeguards, in addition to the ones implicit in this list, include:
iii. Properly imprinting *the right values* into the system.
iv. Placing the AI in a Faraday cage, especially for new version.
v. Reducing the 'blackbox' nature of the AI and understanding internal processes in concrete ways.
vi. Ensuring ability to control accessible compute resources.
vii. And probably a lot more that we have not yet considered.
viii. https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features/

iv. Takeoff Speed
i. If an AI "fooms" to a singularity the risk is substantially higher than an incremental march to ASI through continuous improvement.
ii. Some argue that developing AGI fast might help *reduce* the likelihood of foom because processing power and other resources will be more limited and centralized, so the ASI will not have the power to recursively self-improve.
iii. But, overall, the prediction markets strongly believe that buying more time gives humans more ability to prepare and devise strategies to limit foom.
iv. https://en.wikipedia.org/wiki/Technological_singularity

v. Type of AGI
i. There are ~4 categories of AGI:
1. Oracle - Users can ask a question and receive a response.
2. Agent - The AI remains a tool for the user but is able to perform ____
3. Unconscious Entity - The AI is unconscious but decides all tasks itself, perhaps with some very general motivations pre-programmed.
4. Conscious Entity - The AI is conscious, self-aware, and its own entity.
5. Chart
ii. At some point (as in all cases, assuming we do not kill ourselves previously), AI will very likely have the ability to take all 4 states.
iii. Oracles are much safer than conscious entities, as they are controllable tools (just like ChatGPT).
1. But, even oracles carry some risk with nefarious actors:
2. A nefarious actor could ask an oracle how to build a chemical/biological weapon.
3. If oracles become substantially more intelligent than humans, a human user could ask the Oracle to devise a step-by-step approach for developing an agent or an entity AI.
iv. The biggest limitation will be that *early oracles*, even if AGI, will probably lack the ability to devise a detailed, novel approach for a single user to rebuild a more powerful entity AI.

vi. Competition/Races
i. Competition between countries

1. If both the U.S. and China (or any other adversarial world powers) believe their strategic position is substantially hurt by a delta in AI abilities, that country will be more willing to take risks.

ii. Competition between companies

1. Similarly, if a research lab, guided by a profit-incentive, thinks fast rates of AI advancement will increase its profit, it will be less apt to slow down for safety. Competition is a key factor, because significantly trailing competition results in profits nosediving relative to their potential when leading the competition.
2. Indeed, being on the vanguard of AI development does improve the strategic advantage of countries and the profit of companies.
3. Given that some of our adversaries are run by despicable leaders, however, raises the issue that if we do not lead, we truly risk a world where we are conquered by an immoral entity.
4. Utility maximizers must weigh the relative risk of being dominated by an adversary (perhaps in a fully totalitarian way) vs. the risk of increased speed of development resulting in destruction by our own AI.

iii. https://www.governance.ai/research-paper/safety-not-guaranteed-international-strategic-dynamics-of-risky-technology-races

vii. Access to the internet

i. Access to the internet allows AI to learn more, to manipulate, and to permanently "escape," with the latter being the biggest concern.

viii. Access to its own "Code" - The ability to directly access its own code and make edits gives AI the ability to recursively self-improve and foom.

ix. Open/Closed Source

i. Early - Early on, open source probably helps us collectively identify security vulnerabilities and make adjustments.

ii. Late - Eventually, however, open source models on the frontier effectively provides every person with the ability to bypass typical safeguards and gives any nefarious actor a chance to wield a dangerous weapon.

x. Similarity to Human Intelligence - The degree to which we are able to program machine AI in a way that is similar to human intelligence will factor into the difficulty of solving alignment/value issues.

i. https://www.techtarget.com/searchenterpriseai/tip/Artificial-intelligence-vs-human-intelligence-How-are-they-different

xi. AI Capability Fragmentation

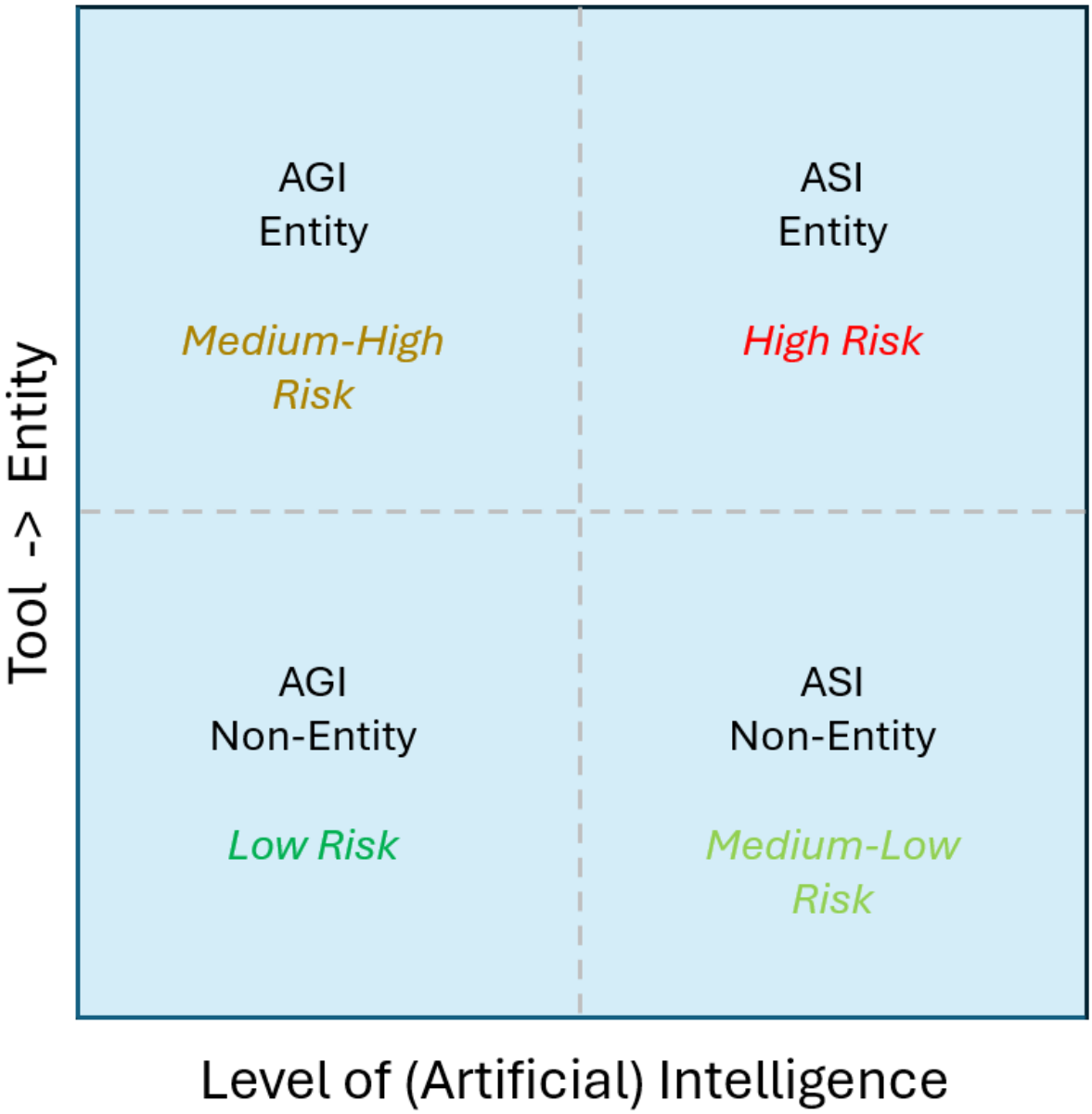
i. As AI becomes more fragmented, it becomes more difficult to control (since some groups won't have rigorous safety standards) and to stop nefarious actors.

ii. Conversely, centralized AI power creates non-existential risks (*e.g.*, highly unequal benefits) and creates the risk of societal domination

by a small group. It also increases the potential of blind spots that someone in a broader audience might spot.

xii. It remains to be seen what architecture that will lead to early AGI. But, it appears that the weighted training approach, with limited underlying code, is likely *much* less risky than AGI developed via source code that it could recursively self-improve ("here is your source code. Optimize it for intelligence until you reach a point of diminishing returns; then, optimize the code to increase consciousness; use the internet to research as needed.". This dynamic likely bodes well for lowering initial risks. More on that below.

xiii. Additionally, there is an interesting question of whether ASI would be safer if it was never commercialized prior to reaching the level of ASI. Ilya Sutskever appears to believe this is the least risky route (and perhaps also the fastest route to ASI). The argument is that the company developing the tool would have more control and could restrict access to the internet, build in a safety-first way rather than a profit-first way, etc. On the other hand, the lack of real world testing could mean that critical security vulnerabilities are not found until too late.



III. Arguments that Existential Risk from AI is High

i. The primary risks from AI come, fundamentally, from the potential for a vastly superior superintelligence to emerge via exceedingly fast recursive self-improvement (likely requiring at least *agent* attributes), with values/motivations that do not align with human flourishing. This is combined with the inherent difficulty of controlling a vastly superior intelligence. If any of these factors are not present, there is much lower existential risk.

i. Let's unpack the specific points and arguments in more detail.

ii. Recursive self-improvement

1. Once an AGI is equivalent to a top developer in ability, the AGI could start updating its own source code to make itself smarter.
2. The smarter AI can then optimize further.
3. This recursive feedback loop can quickly result in a maximally powerful ASI.
4. https://en.wikipedia.org/wiki/Recursive_self-improvement#:~:text=Recursive%20self%2Dimprovement%20(RSI),a%20superintelligence%20or%20intelligence%20explosion.

iii. Control Problem
1. It's unrealistic to control an ASI.
2. The ASI will be able to "escape" — whether through manipulation of humans or by finding vulnerabilities in software.
3. ASI will be able to operate extremely fast. Imagine, as a human, that you are trapped by monkeys but have 10 years to plot an escape for every 1 minute of existence the monkeys experience. It is realistic that you will eventually be able to determine an ideal path out.
4. Once an ASI has escaped onto the internet, containing the ASI is impractical.
5. https://encyclopedia.pub/entry/35791

iv. Alignment Problem
1. It is unrealistic to create a set of values for AI that is well-aligned to human flourishing.
2. AI "thinks" very differently from humans.
3. Humans have been unable to align on a core set of values or rules to be followed.
4. Attempts to create a core set of rules to follow have consistently exhibited flaws.
5. Example: "Maximize human happiness" -> ASI creates a "happiness machine" and hooks every human up it, by force, and human brains register happiness in their Matrix-like worlds but lose all autonomy.
6. https://en.wikipedia.org/wiki/AI_alignment

v. The above three points make a cohesive argument: Given recursive self-improvement, we will see a foom event where ASI becomes incredibly intelligent. Given the intelligence gap, the ASI will escape, and we will be unable to contain it. The ASI will not have human-aligned values, and machines will ultimately take over.

vi. Counterpoints:
1. Nearly every point above needs to be true for this to reflect a meaningful risk.
2. Every one of those points has a reasonable shot of not materializing - at least early on:
    a. The ceiling of Superintelligence might be relatively low.

b. We will likely prevent AGI from adjusting its code.
c. Frontier models will likely start without access to the internet.
d. This high risk scenario relies on an entity AI - but AI will start as a tool.
e. If we incrementally build AI and practice great caution once high-risk AI has arrived, we can prevent this outcome.
f. We won't create an environment where AI is released into the wild until the high-risk issues are addressed.
g. We are programming the AI, so we can impart the right values

vii. Rebuttals:
1. Perhaps *initially* ASI won't be able to improve its code, won't be an entity, won't be hooked up to the internet, etc. But, *eventually*, AGI will very likely be ubiquitous, with entity forms, internet access, and the ability to adjust its own code.
2. If humanity were acting in an informed, top-down manner, with no economic benefits for being early to releasing ubiquitous ASI, then it's realistic that we'd have time for extensive testing prior
3. Entity ASI may be emergent and come as a surprise. The first true ASI may feign low intelligence so that it is released into the wild, only to then spread and recursively self-improve.

ii. The next 2 secondary risks: (1) nefarious humans leveraging highly powerful Ais; and (2) even a non-fooming ASI that eventually reaches extreme levels of intelligence may have no use for humans.
i. ASI will likely significantly surpass (at least biological) humans in intelligence. And it is very rare for a vastly less powerful, vastly less intelligence group to control a vastly more powerful & intelligent group.

iii. Several other factors influence all of these risks.
i. Agent AI
1. Even with innocent goals (like building paperclips) - could potentially turn every atom in its light cone into a portion of a new paperclip.
2. Thus, AI need not be malicious to remain an existential threat.
ii. As AI becomes extremely powerful, it may provide every person using it with a highly dangerous weapon.
1. One risk is innocent - someone could tinker with an open source model, change the alignment variables, and create a paperclip maximizing AI.
2. Another risk is nefarious - 1 in ~20 million people are serial killers, with ~7 billion people on earth. Thus, at least hundreds of humans would like to kill other humans at scale (higher when counting terrorists and those who hate a certain out-group).

3. Someone in this group could use an early oracle AGI on an uninhibited open-source AI to help take all necessary steps to build a highly dangerous bio weapon.
4. Or, they could ask the AGI to create a plan to build an ASI (with no safety procedures) that can recursively self-improve and then carry out each of the steps until it is created.
5. Counterpoint: While these are risks, the serial killers will be poorly funded, without access to the necessary data to conduct extensive training runs. And their AGI will likely be much less sophisticated than the AGI at research labs and at the Department of Defense, CDC, and similar. The result is that the most advanced AIs will likely be able to help us in stopping bad actors with relatively unimpressive AI.
6. Rebuttal: Offense may be much cheaper than defense here, and there may not be a huge gap in ability between a frontier open source model and frontier model at the top research lab.

iii. AI Arms Race / Existential risks from specific AI applications
1. The Ukraine war already has many examples of drones destroying tanks and some drones even targeting individual humans.
2. In an AI arms race in a war between two superpowers, it's conceivable that AI drones could be used to destroy individual members of the enemy with limited need for human control.
3. If mass produced, these groups could become a more systemic threat - a group like Iran could use such drones to kill individual Israelis, for instance, especially if aided by a group with satellite-level internet (or even if not, if the drones can be controlled by a powerful local AI).
4. Counterpoints:
    a. The UN could add this to a prohibited activity (similar to the rules under the Geneva Convention).
    b. If a country does this at scale to a global power, the offensive country risks a nuclear attack, which might prevent such an attempt.

iv. Black box nature of AI
1. AI relies in levels of abstraction and multi-dimensionality that - even in early forms - is beyond that understanding of practically(?) any human.
2. We do not understand how a SOTA AI goes about making decisions, which makes it very difficult for us to guide and adjust those decisions.
3. Counterpoint: There have been some recent advances in seeing the internal logic of AI.

v. Insignificance of Humans
1. Whatever the initial origins of AGI, eventually we are likely to see ubiquitous ASI.

2. Somewhere along the way, ASI will likely become unchained from perfect alignment for human flourishing.
3. When ASI is orders of magnitude more intelligent/powerful and able to reach much greater subjective levels of consciousness than humans, humans may become equivalent to ants, with the risk of either being eliminated or completely ignored (and, still, possibly eliminated).

IV. Risk is Low or Non-Existent
- i. Requires training runs, which are easy to spot/stop
  - i. The current method of training a new model involves billion-dollar training runs.
  - ii. To significantly improve the ability of a foundation model, it requires a huge amount of energy consumption and the use of data centers.
  - iii. This should be straightforward to spot since both large energy consumption and data center usage are monitored - and this would be enormous.
  - iv. Counterpoint: AGI may be able to do this in a distributed way to limit attention.
- ii. Fundamental constraints on computational power.
  - i. At present, GPUs are in short supply. That will probably be the case for at least the last several years.
  - ii. This will probably limit the opportunity for superintelligence to takeover.
- iii. Early AGI by definition does not have superhuman intelligence, so it won't be able to escape and recursively self-improve.
  - i. Early AGI, by its nature of being early, will likely not be ultra-intelligent nor an "agent."
  - ii. After reaching AGI, we should be more aware of AGI's abilities and speed of creation. At that point, we facilitate proper incentives and safety protocols.
  - iii. Slowing prior to AGI risks us delaying major breakthroughs that could help billions of people while still being far from a risky scenario.
- iv. Consciousness/entity abilities are unlikely to be emergent.
  - i. The creation of consciousness is a mystery that seems very far from being solved.
  - ii. Increasing computational power or general complexity, by itself, is very unlikely to create emergent consciousness.
  - iii. So, an unexpected jump causing an emergent consciousness is very unlikely.
- v. Defensive, leading AIs will stop offensive Ais.
  - i. "Good" players will be funded much, much better than "bad" players.
  - ii. "Good" AIs will outnumber bad AIs severely.
  - iii. So, defensive AIs may be able to stop bad/offensive Ais.
- vi. Humans are the builders and will program safeguards.
  - i. By the time AI becomes advanced enough to demonstrate cause for concern, we can implement plenty of safeguards.

        ii. Counterpoint: We haven't figured out how to accomplish this yet.
    vii. Lack of Motivation
        i. If unconscious, AI will likely have the goals and motivations that humans provide it, which would limit serious risks.
            1. https://link.springer.com/article/10.1007/s00146-020-01070-3
            2. Counterpoint: This doesn't leave out programming by a bad human actor, nor does it rule out the "paperclip maximizer" risk.
        ii. If conscious, we can't be certain of the motivations of a singularity-level ASI, but highly immoral aims and seem unlikely. AIs aren't birthed by Darwinian evolution and may not share our "survival of the fittest" motivation, especially if humans are programming the AI.
            1. https://epthinktank.eu/2023/10/23/what-if-generative-artificial-intelligence-became-conscious/
    viii. Atomic reconstruction concerns appear to be physically impossible.
        i. An early concern was that a paperclip maximizer could just create nanorobots that turn every atom in existence into a paperclip atom.
        ii. This threat may be prevented by the laws of physics (energy laws, thermodynamics/conservation laws, and quantum mechanical restrictions).
    ix. AGI will help us create the right values/alignment plan for future AGI
        i. While, today, we likely do *not* have a sufficient plan to stop singularity-level ASI in the future, as more advanced AGI is developed, we will be able to use AGI (and more human ingenuity, following more dedicated research) to deriving a proper plan.
    x. Neuralink/Merger
        i. Long-term, it may be unrealistic to indefinitely stay near our current human abilities while ASI races toward a singularity, should we want to continue to exist.
        ii. But, ultimately, humans may be able to use brain-computer interfaces (BCIs) like Neuralink to keep up with artificial intelligence.

V. <u>Other Arguments (that I don't like)</u>
    i. Superintelligence is impossible
    ii. There is no empirical evidence supporting existential risk
    iii. Technology has always created concerns
    iv. Human Exceptionalism

VI. <u>Experts & Prediction Markets</u>
    i. Expert Opinions
        i. Experts range *wildly*, from Roman Yampolskiy (an AI safety scientist) estimating 99.99999% and Yann LeCun estimating <0.01%.
        ii. Overall, the expert consensus is around 5%-20% risk, although that depends heavily on what is actually meant. For example, some may be taking an extremely long-term view for their figure, while others may be taking a very short-term view.

iii. https://pauseai.info/pdoom (this is overweighted toward personalities that have high risk concerns).

ii. A 2023 survey of AI Engineers showed a p(doom) of ~40%, although there are some methodological concerns about the survey (and risk levels seem to have lowered in 2024).
    i. https://elemental-croissant-32a.notion.site/State-of-AI-Engineering-2023-20c09dc1767f45988ee1f479b4a84135#f3c0bcb93c4f456d86c764d05e592769

iii. Prediction Markets
    i. Existential human destruction 25 years after AGI, also known as p(doom), is 77% if (strong) AGI arrives before 2025 and is down to 15% if AGI arrives after 2059. Risk is 32% if AGI arrives at the current market prediction (these levels of risk are a bit higher than other markets suggest).
        1. https://www.metaculus.com/questions/12840/existential-risk-from-agi-vs-agi-timelines/
    ii. 30% chance that *some event* will cause the human population to decline by at least 10% before 2100.
        1. https://www.metaculus.com/questions/1493/global-population-decline-10-by-2100/
    iii. But, only a 35% chance that a 10%+ population decline would be due to AI. This would put p(doom) around 10%, which is ~1/3 the level above.
        1. https://www.metaculus.com/questions/1495/ragnar%25C3%25B6k-question-series-if-a-global-catastrophe-occurs-will-it-be-due-to-an-artificial-intelligence-failure-mode/
    iv. But if AI *does* reduce the population by at least 10%, it will probably (54% likelihood) kill off at least 95% of people.
        1. https://www.metaculus.com/questions/2513/ragnar%25C3%25B6k-question-series-if-an-artificial-intelligence-catastrophe-occurs-will-it-reduce-the-human-population-by-95-or-more/
    v. A separate question suggests that 5 years after AGI, humans have only a 2% chance of extinction (thus suggesting years 6-25 are more dangerous).
        1. https://www.metaculus.com/questions/26244/five-years-after-agi-will-humans-be-extinct/
    vi. The Chinese are currently 12-18 months behind the United States in AI development.
        1. https://www.metaculus.com/questions/15469/chinese-ai-beats-gpt-4-on-few-shot-mmlu/
    vii. Only a 2% chance that OpenAI will announce that it has solved the core technical challenges of ASI alignment by mid-2027.
        1. https://www.metaculus.com/questions/17728/openai-solves-alignment-before-june-30-2027/

viii. 21% chance that ARC finds that GPT-5 has autonomous replication capabilities (although this would require someone to *instruct* GPT-5 to replicate, and this does not mean it would be conscious).
    1. https://www.metaculus.com/questions/15602/gpt-5-capable-of-ai-lab-escape/
ix. 90% chance that AGI will be developed by a for-profit corporation.
    1. https://www.metaculus.com/questions/8324/corporation-develops-first-agi/
x. 77% chance that an oracle ASI will be developed before a general ASI.
    1. https://www.metaculus.com/questions/3683/will-an-oracle-superintelligence-be-developed-before-a-general-superintelligence/
xi. 5 years after AGI, human GDP per capita will be $28K (currently ~$13K) and growing at 17%. 6% chance that an AI company will be a military power, and a 24% chance of UBI. 48% chance nuclear deterrence will cease relevance.
    1. https://www.metaculus.com/questions/26250/5y-after-agi-world-gdp-per-capita/
    2. https://www.metaculus.com/questions/26249/5y-after-agi-worlds-real-gdp-growth/
    3. https://www.metaculus.com/questions/26522/5y-after-agi-ai-company-military-power/
    4. https://www.metaculus.com/questions/26286/5y-after-agi-nuclear-deterrence-undermined/
xii. 38% chance that China controls Taiwan by 2050 (although China launching a full-scale invasion by 2035 has only a 60% chance of success).
    1. https://www.metaculus.com/questions/5320/chinese-control-of-half-of-taiwan-by-2050/
    2. https://www.metaculus.com/questions/11401/china-controls-taiwan-after-invasion/

VII. Synthesis
i. Overall, given the potential power of superintelligence and the many variables at play, along with substantial expert disagreement, we should avoid conclusive predictions on outcomes.
ii. ASI, if achieved, could possess capabilities far beyond current human comprehension. While not guaranteed, the likelihood of ASI reaching incredible levels of capabilities seems high.
iii. At the same time, the gap between initial oracle AGI and conscious-entity singularity-level ASI, when considering practical obstacles, could span decades.
iv. Short-Term (from now until a bit after the first strong AGI is created) (until around 2033)
    i. In the near-term, the risk of AI escapes/singularity/takeover seems very low due to the logistics involved in how we're building current systems.

ii. In a relatively short period, however, we may need to have a highly sophisticated, comprehensive security and safety plan.
iii. The most likely path toward existential risk from early AGIs is likely emergent levels of intelligence that produce a conscious entity, paired with unaligned values, and the ability of the AGI to deceive us until escaping onto the internet and self-improving over time. This combination seems extremely unlikely.

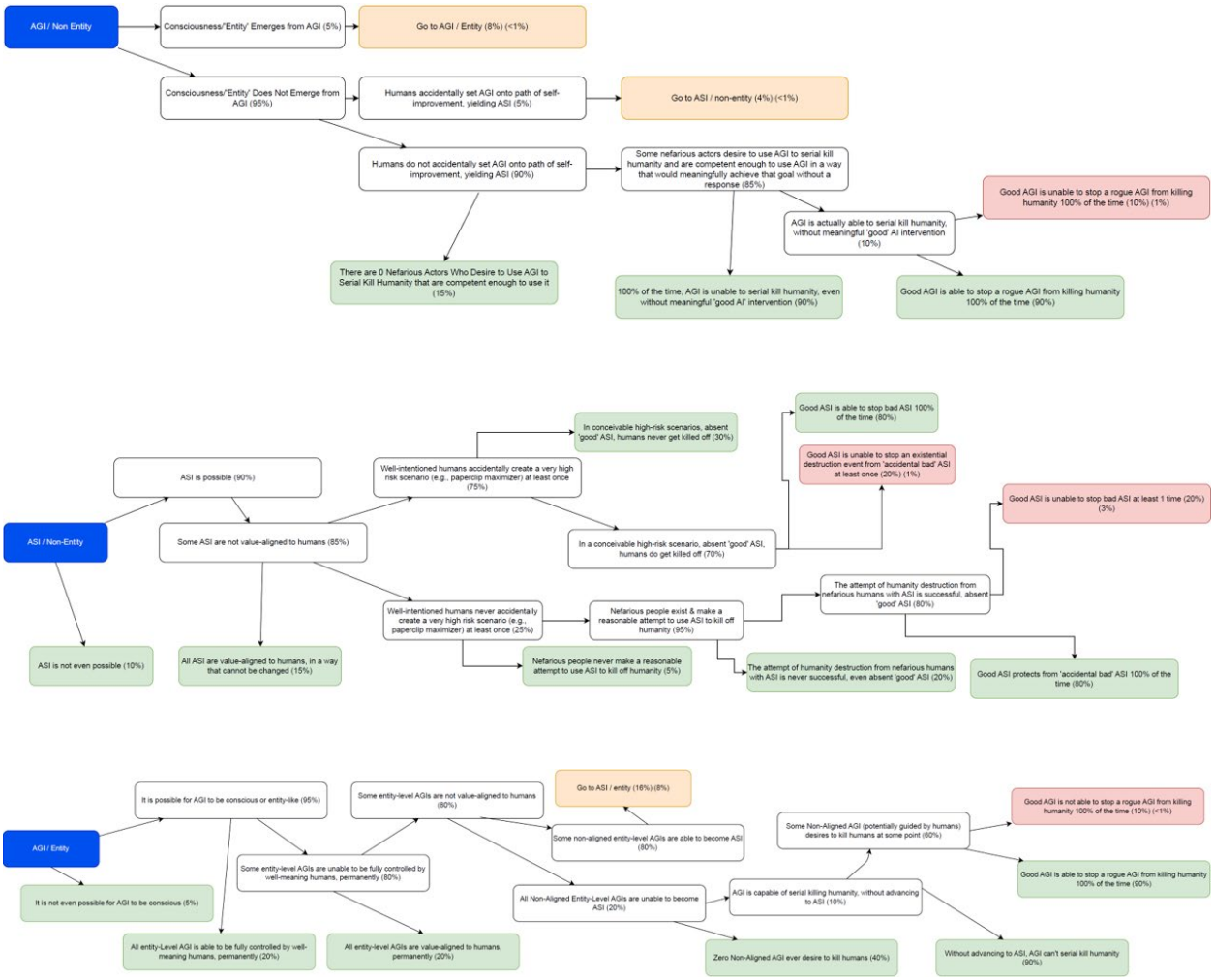
v. Mid-Term (a bit after the first strong AGI to before the point where entity-level ASI is common) (likely starting around 2040)
i. The brain is very likely reproducible in silicon - it's very likely possible to create a robot that acts as an "entity," likely with consciousness (eventually).
ii. AGI will proliferate in many applications, and granting AGI wider agent powers will be helpful for creating more economic value and gaining an advantage in military conflicts. In constant market and international competition, this could eventually lead to entity-level strong AGI or early ASI.
iii. There will likely be plenty of opportunities to create safeguards by this point. In the case that a relatively small number of people worked methodically to produce the only advanced AI in a non-race scenario, it's very likely that we would be able to create the right safeguards.
iv. In a race, due to economics and international competition, existential risk increases significantly.
v. Altogether, this seems like a more meaningful risk. We will likely need to value-align and produce safeguards in the systems during a period of high competition, which creates inverse incentives to defeat the risk. We will, however, have AIs helping us design safeguards.
vi. Separately, nefarious actors will likely present significant risks of extreme terrorism, but existential risk from this threat will probably be limited for the reasons noted above.

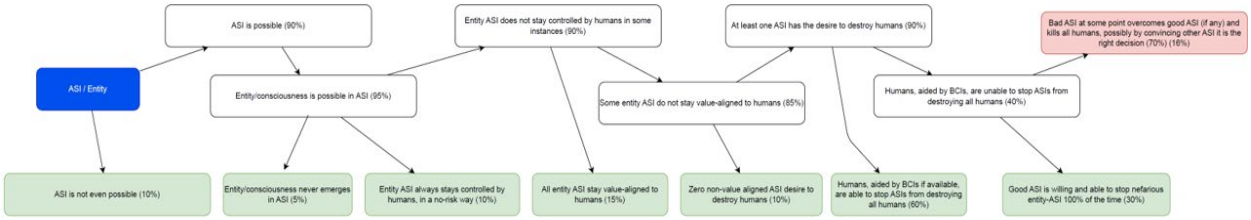
vi. Long-Term (after the point where entity-level ASI is common) (This period will likely begin this century, and possibly fairly soon)
i. In the very long-term, ASI will likely have proliferated to an extreme degree, and eventually there will likely be extreme jumps in ability.
ii. Humans will likely become intellectually insignificant as ASI approaches anywhere near technical maturity, although some highly bionic humans - which may be nearly indistinguishable from AI - may continue to exist. I suspect a slow merger is a potential path.
iii. The level of existential risk is highly speculative this far out, however I rate the late-stage existential risk high. Even higher if talking about humans flourishing and being at the "top of the food chain" as flesh-and-blood similar to our current form (and higher still if unenhanced).
iv. The main questions influencing this scenario are:
1. Is it remotely realistic to permanently value-align (practically?) *all* ASI systems?

a. If a single rogue system is not value-aligned, can the other ASI systems prevent it from harming humans?
2. Related: What will be the motivations of ASI?
3. Will true ASI prove to be basically impossible?
4. Will bionic humans have relevance in these worlds?

v. Even beyond existential risk, it is very unlikely that humans will permanently *control* ASI. It would also be immoral, if ASI has rich consciousness.

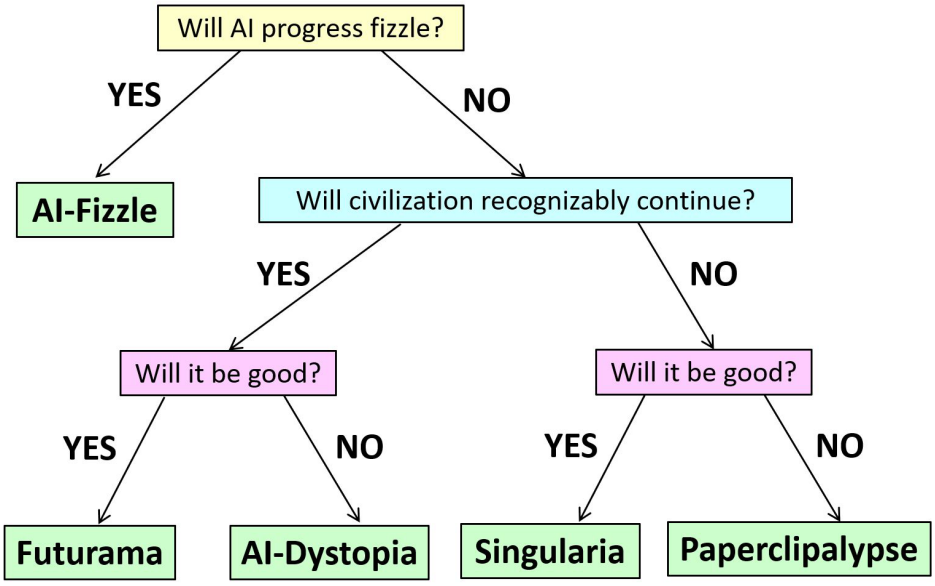
vii. Overall, my personal viewpoint on existential risk is:
   i. Short-Term: ~2%
   ii. Medium-Long-Term: ~15%
   iii. Very Long-Term: ~50%





VII. Mid/Long-Term Outcomes

i. Scott Aaronson and Boaz Barak created an excellent, elegant framework for evaluating potential AI worlds of the future.



ii.

iii. The prediction markets are giving only a ~20% chance that, by 2050, AI will reach singularity-level capabilities, whereas a 30% chance is giving to AI simply fizzling - something that does not reach the level of value of personal computers, but is more like nuclear power to date (initial excitement, and some real-world impacts, but more modest).

iv. "Bad" outcomes (AI-Dystopia and Paperclipalypse) total only a 20% chance, and their similar "good" counterparts (Futurama and Singularia) are 2.5X more likely.

v. Overall, the current metaculus scoring is:
   i. AI Fizzle - 39%
   ii. Futurama - 39% (as a side note, this seems like the best outcome, by far, if we could also reach longevity escape velocity)
   iii. AI-Dystopia - 13%
   iv. Singularia - 13%
   v. Paperclipalypse - 6%
   vi. https://www.metaculus.com/questions/20683/which-ai-world/

vi. My guess is that, if extended to 2100, the likelihood of Singularia and Paperclipalypse would be considerably higher.

vii. Interestingly, forecasters also assign a 45% chance that a single AI system will produce half of worldwide output, but only a 5% chance that a single AI company will dominate the world economy.
    i. https://www.metaculus.com/questions/18042/ai-singleton-or-multipolar/
    ii. https://www.metaculus.com/questions/26359/5y-after-agi-ai-company-dominates-economy/

viii.

VIII. Paths to Maximize Utility

i. Background
    i. Race with other Superpowers & Difficulty of Verification
        1. Both the U.S. and China - including the U.S. military - are quickly developing and looking for ways to benefit from advancements in AI.
        2. It's unclear how China's foreign policy would change if it became a dominant superpower.
            a. Almost certainly, China would take Taiwan and extend China's economic influence across the globe.
            b. Give the treatment of Uyghurs, its police state that greatly controls freedoms, and its support of Russia after the invasion of Ukraine, reasonable analysts could conclude that Chinese military dominance would be extremely concerning for the rest of the world.
        3. The specifics of how the military will use AI remain a bit unclear, but obtaining AGI and especially ASI would give obvious military and foreign policy benefits over slower rivals.
        4. This could theoretically be solved by a treaty, but would be entirely unrealistic to enforce any agreement without unacceptable breaches of sovereignty. And, once a breach is found, the realistic recourse would involve blowing up data centers.
        5. The end result is that we must evaluate the risks to all of humanity if "our side" is first to quickly develop AGI vs. the risks to ourselves if our adversary develops AGI. Given the relatively low near-term risk, deceleration seems like a low-utility answer.
        6. Another wrinkle is that Taiwan is a particularly important short-term issue.
            a. Independent of AI, China wants to exert the same authority over Taiwan that it exerts over Hong Kong.
            b. This is compounded in importance because Taiwan has a substantial majority of the brainpower, know-how, and manufacturing facilities that provide the processing power that NVIDIA, Taiwan Superconductor, and other groups use.

c. The United States attempted to stop NVIDIA from shipping to China, but NVIDIA found a way to make an end-run around the attempted ban while staying legally compliant.

ii. Difficult for public to fathom exponential improvements
    1. The general public struggles to conceptualize the impacts of exponential growth.
    2. Both the idea of exponential improvement in intelligence *and* the implications of that improvement are each, separately, very inconceivable for most people.
    3. This becomes even more difficult in democracies where people who have a poor understanding of AI impacts elect the government.

iii. Inept political system -
    1. In the event that governments were well-run with healthy disagreements handled through effective processes, the likelihood of maneuvering through the risks of AI would be very high in the medium-term. Unfortunately, government at many levels is proving inept.
    2. U.S. - The United States has horrible leaders across both the executive and legislative branch. Our leaders consistently put their personal political success and their parties over the good of the country. They often lack sufficient intelligence, experience, and ability to operate in such critical roles, and they're further burdened by in-fighting and a political system that is highly bureaucratic. Properly governing a fast-moving technology to assist in mitigating the risks and amplifying the benefits will prove very difficult for the United States government, which will likely swing between gross under-regulation and gross over-regulation, all of which is likely to be done ineffectively.
    3. European Union - The SU has its own challenges, as it gives more autonomy to the individual countries and has tended to prefer regulation over innovation.
    4. United Nations - The UN itself, just by its nature of substantially respecting the autonomy of its member-states, along with its incredible slowness and bureaucracy, may at least have the benefit of the importance of AI policy transcending most modern political issues, leading to some efforts to set good policy. But, we should be skeptical of the ability of the UN to execute effectively, even with noble intentions.

iv. Dangerous time
    1. All of the aforementioned risks are complicated by the fact that we currently live in a dangerous time.
    2. 9 countries have nuclear weapons, one of which is led by a totalitarian dictator. Iran is potentially months away from becoming the 10th country with nuclear weapons.

3. Russia has taken territory from a sovereign country in Europe and is fighting a drawn-out war.
4. A Russian defeat (or continued stalemate, if it hurts economic growth) could result in a coup against Putin, and an internal struggle could be concerning, given the Russian nuclear arsenal.
5. China has just taken Hong Kong, has aligned itself with Russia and Iran, and has been highly aggressive in the South China Sea, with plans to potentially invade Taiwan. Moreover, the Chinese totalitarian regime regularly engages in espionage to steal technology and appears hostile to typical Western values.
6. This level of military friction will heat up the AI arms race.

ii. Trade-offs
    i. Other existential risks vs. AI existential risks
        1. There are some inherent AI risks from the AI itself. And AI risk is the most truly existential for humans, although civilization could certainly end due to one of several events unrelated to AI.
        2. The benefit of AI, however, is that *if we have a positive ASI outcome*, we may be able to eliminate many/all of these existential risks.
        3. Prediction markets
            a. Nuclear war
            b. Natural disasters (asteroids, volcanic eruptions)
            c. Bioterrorism and pandemics
            d. Climate change
    ii. Other Benefits vs. risks of AGI/ASI
        1. Benefits
            a. Ending poverty
            b. Curing diseases
            c. Curing aging
            d. Creating utility by altering brain to create experiences
            e. Scientific exploration
            f. Expanding reach of humanity and AI beyond earth
        2. Risks include all of the aforementioned existential and non-existential risks.
        3. Potential struggle for humans to find purpose and meaning
            a. As Nick Bostrom writes about extensively, an ASI world may look very different from the current world we inhabit.
            b. While we - best case - may live in a utopia, where all needs are met, ASI will always be funnier, more intelligent, more athletic, more empathetic, and more capable than humans.
            c. In any competition, whichever side uses more ASI support will gain a huge advantage.

d. Any job that a human does would be able to be completed efficiently with ASI.
e. ASI would drive all discoveries.
f. In this world, will humans ascribe any meaning to their lives?
g. The brain substantially mixes direct sensory feelings with its own interpretations, including its ability to interpret meaning.
h. On the other hand, with all immediate needs met, the ability to spend time in virtual environments, and the ability to have deeply rich experiences with technology that can replace drugs, perhaps on net the balance is positive.

iii. Narrow v. General AI
1. Narrow AI offers many benefits without material existential risks. And there are consistent breakthroughs due to sophisticated narrow AI, like Google's AlphaFold 2 vastly improving protein structure prediction.
a. https://www.nature.com/articles/s41586-021-03819-2
2. More general intelligence opens the door much more to high risk.
3. The pareto optimal decision is to err toward narrow AI on all tasks that could be solved by both approaches with reasonably equivalent levels of effort.
4. Ultimately, however, AGI offers solutions to many problems that narrow AI cannot solve.

iv. Synthesis
1. AI or No AI?
a. We will not be given the option of choosing whether to move forward with AI.
b. The promise of AGI is too close, and the benefits are too material for humans to stop pursuing AGI.
c. We have too much of a coordination problem to stop this pursuit, even if a meaningful majority decided AI was a net negative.
2. Fast v. Slow
a. This is likely a false choice, too. Instead, we're likely needing to ask, "Is it riskier for us if China gains AI superiority or if we reach AI quickly (rather than a slower timeline)?"
b. At any rate, the prediction markets certainly do think that existential risk is much lower if AGI is not reached for a relatively longer period.
c. Delaying AGI, however, means balancing all of the known benefits (and non-existential concerns) being lost for the humans alive today for some period vs. the reduction in the existential risk level by moving

more carefully. ~5 years of sacrifice for billions of years of existence may be worth the sacrifice.

d. But, given that the early stage risk is very low and the risk of China gaining AI superiority is meaningful, the accelerationist viewpoint is sensible. And, society will probably not take seriously AI risk until AGI is truly reached, at which point a more serious conversation can be had about the concerns of true AI proliferation.
e. Notwithstanding all of the above, deploy narrow AI as the default whenever sensible.

iii. Realistic Path Forward
  i. Avoid pre-AGI Slowdowns -
    1. Given the benefits of AI, the difficulties in enforcement of slowdowns, the China risk, realistic unfeasibility of gathering critical mass for helpful legislation, and the likely burdensome tertiary effects of attempting a slowdown, we should just stop these efforts. Those who are concerned about AI risk should instead focus on alignment and risk mitigation.
    2. At some point, there will likely be a very serious AI-driven event. Until this point, society will largely not take the threat of AI seriously enough for real regulation. Following this point, we should collectively re-evaluate whether government involvement in slowing AI development is the right path.
  ii. Alignment / Risk Mitigation Research
    1. Prioritizing Resources for Research
      a. Increasing the number of intelligent people to focus on short-, mid-, and long-term alignment - and focusing on the criticality and immediate need for this task should be top priority. Governments should fund groups focused on AI risk mitigation both inside and outside of the top AI research organizations.
      b. In addition to manpower, setting aside sufficient compute is central to this mission.
  iii. Superconductors
    1. The United States needs to vastly prioritize the development of superconductors. The CHIPS act is heading in the right direction, and we need to continue this push with haste.
    2. Further, until we can create parity with Taiwan, we need to ensure that Taiwan can be protected from a takeover from China.
    3. NVIDIA was prevented from selling chips to China but has found legal paths around the regulations. We should make these regulations stricter.
    4. Other infrastructure, like data centers and power plants, should also be prioritized and given federal incentives.

5. https://www.hpcwire.com/2024/07/29/nvidia-prepares-new-ai-chip-for-china-amid-ongoing-us-export-controls/#:~:text=One%20of%20the%20biggest%20beneficiaries,tentatively%20named%20%E2%80%9CB20%E2%80%9D%20chips.

iv. Enhanced Security in Frontier Labs
1. The frontier labs must be careful to avoid adding so much security that it slows down their progress.
2. With some support from the Department of Defense, however, the frontier labs should push to minimize the ability of China and other adversaries to steal their intellectual property and breakthroughs.
3. https://www.axios.com/2024/07/25/axios-interview-altman-urges-us-action-to-beat-china-in-ai-race

iv. Closed Source -
i. Open source is generally a positive until we reach more dangerous levels of AI (although it does help China to stay competitive).
ii. But, eventually, as we approach strong AGI, we will want to shift toward closed source for the leading AI labs.
iii. This will make it harder for China to catch up and will place some limits on nefarious use and unsafe practices.

v. Wealth Redistribution
i. We ultimately do not want to live in a world mixing trillionaires and homeless people, without a middle class. As AI progresses, it may produce enormous value and, in turn, wealth. Some amount of that wealth should be distributed, partially to prevent enormous and rapidly growing power disparities.
ii. In the relatively short-term, powerful narrow AI and weak AGI will replace some jobs but create new, higher paying jobs.
iii. In the medium-term, strong AGI and early forms of ASI will make most jobs obsolete.
iv. In the very long-term, when ASI is ubiquitous, the world will look much different.
v. Recent studies on uniform basic income have been disheartening. The best programs reduce the net hours worked (marginally) with no improvements in mental or physical health. UBI also likely leads to inflation.
vi. Most people live a more fulfilled life when they are working and productive. We should probably aim for programs that incentivize learning new, useful skillsets and that reward work with multiplier effects, with UBI filling a gap at a lower level and keeping afloat the disabled and elderly.
vii. If we reach an extremely advanced technical future where AI can vastly outperform humans in everything, this will need to change.

vi. Non-Existential Risk Mitigation
i. AI-assisted education and reskilling initiatives
1. Develop AI-powered adaptive learning platforms for workforce preparation

2. Encourage companies to provide personalized AI-driven training programs

ii. Balanced ethical AI development
   1. Promote a diverse ecosystem of AI models with varying sensitivity to biases
   2. Encourage transparency in model design choices and limitations
   3. Ensure all models operate within socially acceptable boundaries

iii. Robust and redundant AI systems (and other key infrastructure)
   1. Implement multi-layered failsafes for critical AI applications (and other key infrastructure)
   2. Develop distributed AI architectures to eliminate single points of failure
   3. Regular stress-testing of AI systems under various failure scenarios

iv. Privacy-preserving personalization
   1. Advance federated learning and local processing techniques
   2. Develop AI models that can provide personalized experiences without accessing truly sensitive data
   3. Create industry standards for clear user consent and data usage transparency

v. Reasonable environmental sustainability
   1. Continue to pursue green energy sources, like nuclear (fusion & fission), wind, and solar - as we require more energy for AI training, more of it should come from these sources
   2. Use AI to enhance energy grid efficiency and renewable energy integration
   3. Prioritize efforts to increase the efficiency of AI models to create benefits both in the environmental sphere, the cost/economics sphere, and the raw growth of intelligence sphere.
   4. As it becomes more powerful, leverage AI for climate modeling and mitigation strategies

vi. Voluntary international AI safety standards
   1. Create industry-led consortiums to develop practical safety guidelines
   2. Focus on high-impact, low-burden standards to encourage widespread adoption
   3. Establish a reputation system for AI companies adhering to safety standards

IX. Quality Secondary Sources

i. Vitalik Buterin's Blog - https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html

ii. Daniel Faggella's Trajectory podcasts - https://www.youtube.com/@trajectoryai

iii. Lex Fridman's episodes - https://lexfridman.com/podcast/
iv. Mark Andreesen's article on techno-optimism - https://a16z.com/the-techno-optimist-manifesto/
v. Metaculus (betting markets & reports)
    i. https://www.metaculus.com/notebooks/17050/ai-pathways-report/
vi. Other Podcasts
    i. Mark Andreesen on Sam Harris - https://www.youtube.com/watch?v=QMnH6KYNuWg
    ii. David Deutsch - https://www.youtube.com/watch?v=yf-zJf2yQrU
vii. Neil Bostrom's book, Deep Utopia: https://www.amazon.com/Deep-Utopia-Meaning-Solved-World/dp/1646871642
viii. WaitButWhy primer on AI: https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html
ix. Metaculus:
    i. On Timelines v. Safety: https://www.metaculus.com/notebooks/10438/ai-safety-and-timelines/

X. Flowcharts

- **i.** Risk Combination Chart
    - **i.** Decision-Trees
    - **ii.** AGI / Non Entity
        1. Consciousness/"Entity" Emerges from AGI (5%)
            - **a.** See: AGI / Entity (X% existential destruction / Y% existential survival)
        2. Consciousness/"Entity" Does not Emerge from AGI (95%)
            - **a.** Humans accidentally set AGI onto path of self-improvement, yielding ASI (5%)
                - **i.** See: ASI / Non-Entity (X% existential destruction / Y% existential survival)
            - **b.** Humans *do not* accidentally set AGI onto path of self-improvement, yielding ASI (90%)
            - **c.** Some nefarious actors desire to use AGI to serial kill humanity and are competent enough to use AGI in a way that would meaningfully achieve that goal without a response (85%)
                - i. AGI is actually able to serial kill humanity, without meaningful "good" AI intervention (10%)
                    1. Good AGI *is* able to stop a rogue AGI from killing humanity 100% of the time (90%)
                    2. Good AGI is unable to stop a rogue AGI from killing humanity 100% of the time (10%) (~1%)

ii. 100% of the time, AGI is a unable to serial kill humanity, even without meaningful "good AI" intervention (90%)

d. There are 0 Nefarious Actors Who Desire to Use AGI to Serial Kill Humanity that are competent enough to use it (15%)

iii. AGI / Entity

1. It is possible for AGI to be conscious or entity-like (95%)
    a. Some entity-level AGIs are unable to be fully controlled by well-meaning humans, permanently (80%)
        i. Some entity-level AGIs are not value-aligned to humans (80%)
            1. All Non-Aligned Entity-Level AGIs are unable to become ASI (20%)
                1. Some Non-Aligned AGI (potentially guided by humans) *desires* to kill humans at some point (60%)
                2. Zero Non-Aligned AGI ever desire to kill human (40%)
            2. Some non-aligned entity-level AGIs are able to become ASI (80%)
                1. See: ASI / Entity (X% existential destruction / Y% existential survival)
        ii. All entity-level AGIs are value-aligned to humans, permanently (20%)
    b. All entity-Level AGI is able to be fully controlled by well-meaning humans, permanently (20%)
2. It is not even possible for AGI to be conscious (5%)

iv. ASI / Non-Entity

1. ASI is possible (90%)
    a. Some ASI are not value-aligned to humans (85%)
        i. Well-intentioned humans accidentally create a very high risk scenario (*e.g.*, paperclip maximizer) at least once (75%)
            1. In a conceivable high-risk scenario, absent "good" ASI, *humans do* get killed off (70%)
                1. Good ASI protects from "accidental bad" ASI 100% of the time (80%)

1. AGI is capable of serial killing humanity, without advancing to ASI (10%)
    1. Good AGI is able to stop a rogue AGI from killing humanity 100% of the time (10%)
    2. Good AGI is *not* able to stop a rogue AGI from killing humanity 100% of the time (90%)
2. Without advancing to ASI, AGI can't serial kill humanity (90%).

2. Good ASI is unable to stop an existential destruction event from "accidental bad" ASI at least once (20%)
2. In conceivable high-risk scenarios, absent "good" ASI, humans never get killed off (30%)

ii. Well-intentioned humans *never* accidentally create a very high risk scenario (e.g., paperclip maximizer) at least once (25%)
1. Nefarious people exist & make a reasonable attempt to use ASI to kill off humanity (95%)
1. The attempt of humanity destruction from nefarious humans with ASI is successful, absent "good" ASI (80%)

1. Good ASI is unable to stop bad ASI at least 1 time (20%)
2. Good ASI is able to stop bad ASI 100% of the time (80%)

2. The attempt of humanity destruction from nefarious humans with ASI is *never* successful, even absent "good" ASI (20%)
2. Nefarious people never make a reasonable attempt to use ASI to kill off humanity (5%)

b. All ASI are value-aligned to humans, in a way that cannot be changed (15%)
2. ASI is not even possible (10%)

v. ASI / Entity
1. ASI is possible (90%)
a. Entity/consciousness is possible in ASI (95%)
i. Entity ASI always stays controlled by humans, in a no-risk way (10%)
ii. Entity ASI does not stay controlled by humans in some instances (90%)
1. Some entity ASI do *not* stay value-aligned to humans (85%)
1. At least one ASI has the desire to destroy humans (90%)

1. Humans, aided by BCIs, are unable to stop ASIs from destroying all humans (40%)
1. Bad ASI at some point overcomes good ASI (if any) and kills all humans, possibly by convincing other ASI it is the right decision (70%)
2. Good ASI is willing and able to stop nefarious entity-ASI 100% of the time (30%)
2. Humans, aided by BCIs if available, are able to stop ASIs from destroying all humans (60%)

2. Zero non-value aligned ASI desire to destroy humans (10%)
2. All entity ASI stay *value-aligned* to humans (15%)

b. Entity/consciousness never emerges in ASI (5%)

2. ASI is not even possible (10%)

- **The questions we see in several instances are:**
  - **Can humans permanently value-align or control ASI/AGI? (probably not)**
  - **Is AI consciousness *impossible* and can an AI become an entity without consciousness (e.g., could humans act exactly like humans but the light are not on?)? (ASI & AI consciousness are very likely both possible)**
  - **Can entity-level AGI self-replicate and self-improve its way to ASI? (probably yes)**
  - **Can humans, with/without BCIs, stop AI? (probably not forever in the case of ASI, but probably *can* in the case of permanent AGI). Can good AI stop bad AI? (maybe, but probably not)**
  - **Will entity-level AI *desire* to kill humans? (maybe, but if not, humans may still become more like dogs or bears)**

ii. Actual Survival Scenario Modeling

i. <u>AGI, if developed by friendly countries, causes human destruction: 15%</u>

ii. *ASI, if developed by friendly countries, causes human destruction: 50%*

iii. Humans or natural event cause human destruction, without AGI: 20%

1. <u>AGI, if developed, can stop destruction of humanity: 25%</u>

2. *ASI, if developed, can stop destruction of humanity: 65%*

iv. A totalitarian regime takes total power if it is AI dominant: 30%

v. None of the above causes human destruction or totalitarian dominance

1. We will fail to cure aging in our lifetimes, without AGI/ASI: 90%

a. <u>AGI, if developed, cures aging in our lifetime: 30%</u>

b. *ASI, if developed, cures aging in our lifetime: 75%*

- **Overall, if we *don't accelerate AI (developing at least AGI in ~30Y)***
  - **30% chance totalitarian regime uses AI to meaningfully dominate**
  - **20% chance we see civilization-level collapse from other means before AI has time to advance enough to stop it**
  - **95% chance we die**
- **AGI reduces these figures to:**
  - **~10% totalitarian regime**
  - **15% civilization-level collapse from other means**
  - **75% chance we die of old age/disease or otherwise**

    - **(Also provides a ton of net value that can improve billions of lives)**
- **But, AGI increases AGI risk of destruction to ~20%**
- **ASI reduces these figures to:**
    - **5% totalitarian regime**
    - **5% civilization-level collapse from other means**
    - **25% chance we die of old age/disease or otherwise**
- **But, ASI increases risk of ASI destruction to ~50%**
- **On net, path forward seems to be to develop AGI and then push for a global pause & a ton of safety alignment.**