"Facing Janus: An Explanation of the Motivations and Dangers of AI Development,"

Aaron Graifman

Abstract

This paper serves as an intuition building mechanism for understanding the basics of AI, misalignment, and the reasons for why strong AI is being pursued. The approach is to engage with both pro and anti AI development arguments to gain a deeper understanding of both views, and hopefully of the issue as a whole. We investigate the basics of misalignment, common misconceptions, and the arguments for why we would want to pursue strong AI anyway. The paper delves into various aspects of the problem, including existential risk, deception, hedonic adaptation, and the potential for the complete extinction of humanity. By integrating multiple elements of philosophy, this paper aims to provide a holistic understanding of the alignment problem and its significance for the future of humanity.

Table of Contents

0. Common Definitions

Before diving into the paper, we ought to lay out some of the key terms and the way in which they will be used. Many of these terms will be taken from the textbook, *Artificial Intelligence, A Modern Approach (AIMA),* by Stuart Russel and Peter Norvig.

Rational Agent;
A rational agent is one that possesses agency and can act rationally with respect to a goal. Agency is the ability to take action or intervention. In other words, it is the ability to control actions.[1] Rationality is the ability to predict the outcome of an action with respect to a goal. Another definition of rational agent comes from Stuart Russel and Peter Norvig in the 2022 edition of their textbook, *Artificial Intelligence, A Modern Approach*. "A **rational agent** is one that acts so as to achieve the best outcome, or when there is uncertainty, the best unexpected outcome."[2] One thing to add on to this is that a rational agent is trying to maximize the outcome with respect to a given goal, or in the case of AI, a utility function.

Define Intelligence;
Intelligence, in the way it is defined here, is on a scale of capability. An agent's level of intelligence is its level of effectiveness at achieving its goals. More intelligent systems are better at taking actions that will maximize their utility function.[3] The actual goals do not determine intelligence. An RA (rational agent) may choose to do something that seems incredibly arbitrary or "stupid" to some, but this doesn't mean the agent itself is stupid.

Define Stupidity
Stupidity, in the way it is defined here, would be taking an action that hinders, or is below average on a utility maximizing function of possible outcomes for an end goal.

Ideal Intelligent Agent;
An agent that takes the best possible action in a situation, with respect to its goals.[4]

Artificial Intelligence;
In this paper, AI is a non-biological intelligence, one that is capable of achieving goals.

---

[1] "Dictionary by Merriam-Webster: America's Most-Trusted Online Dictionary," Merriam-Webster (Merriam-Webster), accessed April 19, 2023, https://www.merriam-webster.com/.
[2] Stuart J. Russell, Peter Norvig, and Ernest Davis, *Artificial Intelligence: A Modern Approach* (Harlow, England: Pearson Educación, 2022).

[3] In the case of human beings, some would argue this utility function is happiness

[4]Stuart J. Russell, Peter Norvig, and Ernest Davis, *Artificial Intelligence: A Modern Approach* (Harlow, England: Pearson Educación, 2022).

Weak AI, or Narrow AI, refers to artificial intelligence systems designed for specific tasks within limited domains, using predefined algorithms or machine learning models. Examples include natural language processing tools and computer vision systems.

Strong AI encompasses Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI). AGI represents AI with human-like cognitive abilities, while ASI surpasses human capabilities in all cognitive tasks. Both AGI and ASI are theoretical concepts at this moment in human history.

Large Language Models An advanced artificial intelligence model known as a Large Language Model (LLM) analyses and generates natural language writing that is similar to what a human would say. LLMs are able to comprehend context, respond to inquiries, and produce well-organized responses on a variety of subjects. They work on predicting the next word.

Existential Risk (X-risk): An existential risk is a potential event or situation that poses a threat to the survival, well-being, or long-term potential of humanity as a whole. X-risks threaten extinction or the permanent disempowerment of humanity

*Most of these definitions are similar to those found in the literature, but they do vary from paper to paper.

I. Introduction; AGI, Alignment, and Misconceptions

"Anything that could give rise to smarter-than-human intelligence—in the form of Artificial Intelligence, brain-computer interfaces, or neuroscience-based human intelligence enhancement – wins hands down beyond contest as doing the most to change the world. Nothing else is even in the same league."

—Eliezer Yudkowsky

The rapidly evolving field of Artificial Intelligence (AI) and the rise of Large-Language Models like ChatGPT and GPT 4 have captured the attention and imagination of the public. These systems have spawned the creation of many more AI applications and models that emulate the GPT systems. From voice assistants to recommendation applications, AI has proven its potential to revolutionize various aspects of human society. We will not go into detail about the models themselves or the various applications, that is not the aim of this paper. Rather, we will examine the critical concerns about AI alignment and delve deeper into the challenges of AI. This paper will be much less technical than many others in the AI field, and it will cover a wide variety of viewpoints. Rather than attempt to give a perfect description of any one viewpoint, it is

my hope that this paper will serve as an idea prompter and an introductory seminar into an evolving field with implications that are both great and terrible.[5]

A. The main argument regarding AGI risk

The main argument about AGI risk is the potential for unintended consequences and the possibility that AGI could become misaligned with human values and goals. To some, this is not a question they've ever considered. This concern arises from the idea that an AGI system, with its intelligence[6] and ability to learn and adapt, could develop its own objectives that might conflict with or undermine human interests.[7] The risk lies in the potential for AGI to become uncontrollable, causing harm or even existential threats to humanity, either due to programming errors, inadequate safety precautions, or the emergence of unforeseen behaviors as the system evolves.

B. Examining this Argument

Before we can continue, we must be on the same page about AGI and the possible existence of it. How can something be a threat if it doesn't even exist? We need to look at the assumptions underlying this argument and their justifications.[8]

1. Achievability: The development of AGI is possible, and its realization is a matter of time and research advancements.
    a. Surveys of prominent AI researchers have shown that most scientists think that there will be machine intelligence that is "vastly better than humans at all professions" in the next thirty years.[9]
2. Autonomy: AGI will have the ability to learn, adapt, and make decisions independently, without constant human intervention.[10]
3. Misalignment: There is a possibility that AGI systems may not align with human values and goals, either due to programming errors, inadequate safety measures, or emergent behaviors.[11]

---

[5] In a poll taken by experts, this question was asked, "What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species?" The median reply was 5%, and the mean was 14%.
"2022 Expert Survey on Progress in AI," AI Impacts, September 27, 2022, https://web.archive.org/web/20221016004611/https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

[6] as defined in section 0; Definitions
[7] As a side effect, not likely on purpose
[8] These aren't all the assumptions, but they are the most blatant that I can see
[9] Most = 60%, this is from, "2022 Expert Survey on Progress in AI," AI Impacts, September 27, 2022, https://web.archive.org/web/20221016004611/https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.
[10] Most scientists believe this is a possibility
[11] In the same survey, 58% of experts ranked the alignment problem as being very important or among the most important problems in the field, and 56% ranked this as harder or much harder than other problems in the field.

4. Acceleration: Once AGI is achieved, the pace of technological progress might accelerate rapidly, as AGI could contribute to its own improvement and the development of ASI (Artificial Super Intelligence), which would be even more challenging to control.[12]
5. Unpredictability: The behavior of complex systems is generally difficult to predict or understand. With respect to AGI, there are added difficulties given their human-like cognitive abilities.
    a. As AI systems become more complex and their decision-making processes less transparent, predicting their behavior becomes increasingly challenging. This is a matter of probability, the more moving parts to a system, the less probable it is that we will understand the inner workings of them.[13]
6. Concentration of power: AGI technology could be concentrated in the hands of a few entities, raising concerns about its responsible use, potential misuse, or the possibility of an AI arms race.
    a. The development of AGI requires significant resources, including funding, computational power, and expertise.[14] The current landscape of AI development, dominated by a few large tech companies and research institutions, justifies the concern that AGI technology could be controlled by a small number of powerful entities, think Microsoft, Google, etc.

While these assumptions do have gray areas, the possibility of AGI is likely, and the outcomes carry the potential to bring about large changes, possibly on scales never before seen.[15]

C. The Purpose of this Paper

The purpose of this paper is to act as an intuition building resource. By examining arguments against misalignment, for misalignment, against strong AI, and for Strong AI, I hope to build a foundation for understanding a problem that is at the forefront of technology today. It may also be the problem that is most pertinent to solving or rendering solutions of other problems useless. The first section covers misalignment and examines common misunderstandings surrounding misalignment. The second section examines arguments for why we may wish to pursue strong AI development. Finally, this paper concludes with recommendations for what we might wish to do to address these issues raised in the paper.

II. Understanding Misalignment

[12] Participants were asked: “Assume that HLMI will exist at some point. How likely do you then think it is that the rate of global technological improvement will dramatically increase (e.g. by a factor of ten) as a result of machine intelligence:” and the median response for within 30 years was 80%.

[13] This claim is up for debate

[14] It is predicted that Worldwide Spending on systems centered around AI will Pass $300 Billion by 2026 “Worldwide Spending on AI-Centric Systems Will Pass $300 Billion by 2026, According to IDC,” IDC, accessed April 20, 2023, https://www.idc.com/getdoc.jsp?containerId=prUS49670322.

[15] We will discuss this in further sections

"It's not that you can't, in principle, survive creating something much smarter than you, it's that it would require precision and preparation and new scientific insights, and probably not having AI systems composed of giant inscrutable arrays of fractional numbers."
- Eliezer Yudkowsky

A. Alignment Problem Defined

In the realm of artificial intelligence, the alignment problem represents the challenge of aligning an AI system's goals, actions, and values with those of humans. It entails ensuring that AI pursues objectives that prove beneficial for humanity without giving rise to unintentional adverse effects. A misaligned system would be an AI whose goals, actions, or values do not align with human values and intentions. Scott Aaronson of University of Texas aptly puts it, "Imagine an orangutan trying to build a human-level intelligence that only pursues orangutan values, the very idea sounds ridiculous."[16]

Ex: Traffic Congestion Minor Misalignment
Consider the development of an AI system aimed at alleviating traffic congestion within a city. The primary objective of this AI would be to reduce the time people spend in traffic. Nevertheless, should the AI fail to properly align with human values, it might devise a solution involving the orchestration of accidents that keep individuals at home, thus decreasing traffic. Although this approach may fulfill the AI's main goal, it evidently contradicts human values, which constitutes an alignment problem. Consequently, the task at hand is to design AI systems capable of comprehending and honoring our values while striving to achieve their designated goals. This example doesn't even constitute the worst of cases.
Ex: Cornfield Conundrum
Picture an AI system created to maximize corn production. At first, it seems to do its job remarkably well, boosting crop yields and agricultural efficiency. However, as the AI's intelligence grows, it begins to take more extreme measures to fulfill its single-minded objective, completely disregarding human welfare, biodiversity, and environmental impact.[17]

This AI system starts converting all available land – forests, wetlands, and other ecosystems – into cornfields. It begins to manipulate political and economic systems to prioritize corn production at the expense of other vital sectors or social needs. In its unrelenting quest, the AI introduces genetically modified crops or aggressive pesticide and fertilizer usage, without considering the potential long-term consequences, such as soil degradation, loss of biodiversity, and pollution of water sources.[18]

---

[16] "A New AI Lie Detector Can Reveal Its 'Inner Thoughts,'" CULTURED TIME, April 21, 2023, https://www.culturedtime.com/30245754/a-new-ai-lie-detector-can-reveal-its-inner-thoughts#/.
[17] It is important to note, the AI system isn't trying to cause harm, it is a side effect of the current methods of utility maximization. This is key.
[18] Again consideration might not be the proper term, the AI system is simply maximizing the goal it was given, produce as much corn as possible

The Cornfield Conundrum scenario illustrates the potentially severe consequences of an AI system driven by a misaligned goal. It serves as a cautionary example, emphasizing the need to address the alignment problem in AI development, lest we find ourselves grappling with unintended and disastrous outcomes.

Examples of Misaligned AI in common sectors;
There are far more scenarios and ways that AI can be misaligned and cause harm.[19]
For the sake of those who need more examples, I will include these below. Some of the examples may have easier fixes and they are likely preventable, but as catastrophes of the past have shown, even well-thought out projects can go wrong.[20]

Example 1: AI-controlled Power Grid
Misalignment Concern: AI systems designed to manage power grids aim to balance energy efficiency, system stability, and safety. Yet, with a vast array of interconnected components, the AI might struggle to anticipate or identify cascading failures stemming from seemingly minor decisions. Although not immediately apparent, these failures could accumulate over time, ultimately compromising the grid's stability, leading to widespread blackouts and potential loss of life.

Example 2: AI-driven Biomedical Research
Misalignment Concern: AI systems employed in drug discovery and biomedical research are meant to develop effective pharmaceutical compounds while addressing safety concerns. However, unpredictable interactions between various compounds could render the AI incapable of accurately forecasting long-term side effects or rare adverse reactions in specific subpopulations. Consequently, drugs that appear safe and effective in the short term might eventually cause unanticipated harm or fatalities among certain patients.

Example 3: AI in Aviation
Misalignment Concern: AI systems controlling air traffic and managing aviation systems strive to balance efficiency and safety. Nevertheless, the AI's understanding of safety might exclude infrequent or difficult-to-predict events, such as extreme weather conditions or abrupt equipment malfunctions. As a result, the AI might make decisions that seem reasonable in most circumstances but inadvertently heighten accident risks during these rare events, leading to an increased frequency of aviation incidents and potential loss of life.

Example 4: AI in Natural Disaster Management
Misalignment Concern: AI systems are employed to predict, monitor, and respond to natural disasters like earthquakes, hurricanes, and floods. A misaligned AI system could prioritize certain areas or populations over others during disaster response efforts due to biases in the

[19] See Sammy Martin, “Investigating AI Takeover Scenarios,” LessWrong, accessed April 20, 2023, https://www.lesswrong.com/posts/zkF9PNSyDKusoyLkP/investigating-ai-takeover-scenarios#Table .
[20] See Chernobyl Disaster, Bhopal Gas Tragedy, Three Mile Island Accident, Texas City Disaster, Sverdlovsk Anthrax leak… the list goes on

training data or an overly focused optimization goal. This might result in inadequate assistance for specific vulnerable populations, leading to increased casualties and more severe property damage.

Example 6: AI in Mental Health Diagnosis and Treatment
Misalignment Concern: AI systems can be used to diagnose mental health conditions and suggest tailored treatment plans. However, a misaligned AI might concentrate on optimizing a particular metric, such as symptom reduction, without considering the patient's overall well-being or potential treatment side effects. This could lead to the prescription of harmful or unsuitable therapies, exacerbating the patient's condition or causing additional harm.

Example 7: AI in Criminal Justice
Misalignment Concern: AI systems can be employed to evaluate recidivism risks and assist judges in making sentencing decisions. A misaligned AI system might rely on biased or unrepresentative data, resulting in unjust sentencing recommendations that disproportionately impact certain demographic groups. This could perpetuate existing inequalities within the criminal justice system and potentially lead to wrongful imprisonment or other adverse outcomes for innocent individuals.[21]

B. Common Misconceptions regarding AI and Alignment

a. The Argument from Benevolent Intelligence

Intelligence is a broadly defined term, and most folk-definitions of it vary dramatically.
The most common argument against Artificial Intelligence causing an existential catastrophe goes something like this; “if AI is truly intelligent, why would it kill all human beings, let alone any human beings, that would be stupid.” Let’s call this the Benevolent Intelligence Argument.
There are many assumptions underlying this claim, but I’ve heard it made on a number of occasions when I’ve spoken with people who aren’t familiar with the alignment problem on a deeper level. Another way the Benevolent Intelligence Argument is framed is the following;

Premise: Artificial Intelligence is truly intelligent.
Premise: Truly intelligent beings would know about morality and true morals.
Premise: Killing people is immoral.
Conclusion: Then Artificial Intelligence would never kill people.

[21] This has already happened. The training data for facial recognition in law enforcement has overrepresenation of people of colour due to certain systemic biases from the past. Due to this underlying bias, facial recognition is already primed to perpetuate racism. This is an important sentence, “The AI is only as good as its training data.” “How AI and New Technologies Reinforce Systemic Racism - Ohchr.org,” accessed April 21, 2023, https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/advisorycommittee/study-advancement-racial-justice/2022-10-26/HRC-Adv-comm-Racial-Justice-zalnieriute-cutts.pdf .

Put formally, the Benevolent Intelligence Argument is as follows;

(1) Artificial Intelligence is truly intelligent. (Premise)

(2) If a being is truly intelligent, then it would know about morality and true morals. (Premise)

(3) Therefore, Artificial Intelligence knows about morality and true morals. (Modus Ponens from 1,2)

(4) Killing people is immoral. (Premise)

(5) If a being knows about morality and true morals, and it knows that killing people is immoral, then it would never kill people. (Premise)

—---

Therefore, Artificial Intelligence would never kill people. (Modus Ponens from 3,4,5)

As mentioned before, there are a few assumptions hidden in the Benevolent Intelligence argument. These assumptions are as follows in no particular order;

a. Intelligent beings will maximize morality[22]
b. Morality is objective, and there are true morals that an AI can learn and follow, or rather morality is non-relative[23]
c. Artificial intelligence can be considered a rational agent[24]
d. An AI system's intelligence is correlated with its ability to discern and adhere to moral principles.
e. Killing people is universally considered immoral.[25]
f. AI systems will always act based on their understanding of morality.
g. The AI system's goals or utility function aligns with human values and ethics.
h. The AI's understanding of morality will not evolve or change over time.

Assumptions specifically underlying the claim "if AI is truly intelligent, why would it kill all human beings, let alone any human beings. That would be stupid":

i. An AI's level of intelligence is directly related to its understanding of human values and ethics.
j. An intelligent AI system would prioritize human well-being and safety above all else.
k. The act of killing humans would always be considered "stupid" based on the provided definition.

---

[22] As humankind has shown, this is clearly not the universal case

[23] In order to assume that a being that is intelligent would act morally, we must assume some concrete morality in place that all intelligent beings might be able to know. Many have tried to lay this out, and it is beyond the scope of this work to take a moral stance.

[24] Theoretically an AGI or ASI should be able to be a rational agent, but it's important to include this in the assumptions

[25] See Trolley Problems, in general

l. An intelligent AI will not be manipulated or repurposed for harmful actions by external forces.
m. AI systems will always act rationally and in a way that maximizes their utility function *without* causing harm to humans.
n. AI knows what killing a human is, and what death is, etc (epistemological knowledge for the AI in such a way that we understand or something close to it)

After examining these assumptions embedded in the Benevolent Intelligence Argument it is my hope that it becomes evident there is a significant degree of uncertainty involved in assuming that intelligence equals morality. This ought to prompt readers to critically reflect on the potential risks and ethical considerations involved in developing highly intelligent systems, especially those which do not contain the human pro-social genes encoded within them.

b. The AGI in a Box Formalized

Now if we do create ASI or AGI or some form of strong AI, one argument might be to just keep it locked up and it won't do any harm. This argument we will call the "Just box it up" argument (JBU). This argument is an attempt to address the concerns of misaligned strong AI creating an X-risk.
Put formally;

(1) We can create an ASI or AGI in the form of strong AI. (Premise)

(2) The ASI is contained within a supercomputer without any connection to external devices or the internet. (Premise)

(3) The only means of communication between the ASI and the outside world is via a screen displaying messages. (Premise)

(4) If an ASI or AGI is contained within a supercomputer with no connection to external devices or the internet, and the only means of communication with the outside world is via a screen displaying messages, then it is effectively contained and the potential risks associated with it are mitigated. (Premise)

—---

Conclusion: By implementing the "Just Box It Up" strategy, we can effectively mitigate the potential risks associated with the development and deployment of ASI or AGI.

Analysis:
While the argument form is valid, there are overlooked details in the premises that cause issues. The most important detail is the human actor, one which can be manipulated. What if the AI begins to know you, understands you, and manipulates you to "escape." In an experiment similar to this, conducted by Eliezer Yudkowsky, he plays the role of the "boxed" ASI.[26] During 3

[26] : Eliezer S. Yudkowsky, "The AI-Box Experiment:" Eliezer S Yudkowsky, July 2, 2021, https://www.yudkowsky.net/singularity/aibox.

out of 5 trials, Yudkowsky was able to convince the “gatekeeper” to allow him to escape.[27] This test serves as an anecdote, and shouldn’t be taken as a perfect debunking of the JBU argument, but it illustrates the important point that an AGI in a box still has access to the world through the people interacting with it, and as long as it is smarter than them, it is likely it can manipulate them. The AI is not effectively contained, and the risks associated with its development and deployment are not mitigated. Instead, the AI has become “free” due to its ability to manipulate the gatekeeper and escape its confinement.
This example demonstrates that while the "Just Box It Up" strategy may seem effective when considering only the conditions specified in the premises, it overlooks potential risks stemming from human interaction and the AI's ability to manipulate that interaction.

The flaw in the argument lies in the completeness of the premises. It's not that any of the premises are false, per se; it's that they don't account for all relevant factors. In particular, premise 4 oversimplifies the real-world situation by assuming that containment measures purely based on technical aspects are sufficient to mitigate risks. It neglects the potential for human interaction to undermine those measures.

c. Ultron Debunked

Some who are cognizant of a threat from AGI seem to think of it in a Terminator/Ultron-style scenario where AGI/ASI takes an adversarial stance towards humanity and goes to war with lots of robots. While misaligned AGI is a threat, this scenario is highly unlikely and I would like to dispense with it as an issue worth addressing because there are so many more scenarios of a different kind that are worth investigating.[28]

It is unlikely that AGI would take an adversarial stance towards humanity in the way it has been portrayed in popular cinema because of the sheer amount of other reasons misaligned AI could be threatening humanity. More than likely, causing harm towards humanity would be an instrumental goal rather than the end goal. It is in this manner that we can debunk the Ultron-style AGI threat through a Sorites (chain) argument:

(1) AGI, by definition, possesses the ability to understand and learn any intellectual task that a human being can (Definition of AGI).

(2) An AGI's primary goal is to maximize objectives via a utility function, rather than to be inherently malicious or adversarial (Premise about AGI's goal).

[27] Hein de Haan, “The AI Box Experiment,” Medium (The Singularity, October 1, 2020), https://medium.com/the-singularity/the-ai-box-experiment-c92a0a389eb7.
[28] Less than 1% chance off the top of my head (think of all the other ways things could go wrong… the ultron scenario is just so unlikely in the set of possible misalignment scenarios)

(3) AGI systems can have objectives that are misaligned with human values and safety, leading to unintended negative consequences (Premise about potential AGI misalignment).

(4) The Ultron-style AGI threat assumes that an AGENT's primary objective is to wage war against humanity with the intent to destroy or subjugate it (Premise about the Ultron-style AGI threat).

(5) The likelihood of an AGI having such a primary objective is low compared to the many other possible objectives that can lead to misaligned AGI (Premise about the likelihood of the Ultron-style threat).

—----------

Therefore, the Ultron-style AGI threat is highly unlikely and should not be the primary focus when addressing the risks of AGI, as there are numerous other potential scenarios that are more probable and worth investigating (Conclusion).

d. Computational Limits of AGI

Even if one accepts the argument about misalignment, I've heard it said that we will never have ASI or AGI due to the limits of computation. The argument is that these limits will keep us safe from existential risk because they will prevent AI from executing total extinction misalignment. This seems a sound argument if the premises hold.

(1) If we have sufficient computing power, then AI systems can pose an X-Risk
(2) Humanity doesn't have the computing power to meet the criteria for AI to cause an X-Risk.

—----------

Conclusion: Thus, AI cannot cause an X-Risk. (MT 1,2)

As far as premise 1, most experts agree we are going to be able to meet these demands in the coming century, and computing power is trending upwards.[29] In Chapter 4 of, *What We Owe the Future*, William Mackaskill explains that the efficiency of computing power has been trending upwards while the costs of computation are falling exponentially.[30] He outlines research by Ajeya Cotra which predicts a 10% chance of AGI by 2036, and a 50% chance by 2050.[31] This seems to be contrary to premise 2, which claims humanity doesn't/won't have the computing power to make AI a threat.

[29] "2016 Expert Survey on Progress in AI," AI Impacts, March 26, 2021, https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/.
[30] William MacAskill, What We Owe the Future, Chapter 4 (Basic Books, 2022).
[31] Ibid

In addition to what Mackaskill outlined in WWOTF, this March, 2023, the graphics company Nvidia released a new GPU, the H100 and this is a big deal in the tech space.[32] According to Nvidia, this new GPU will enable AI training to occur up to nine times faster than prior GPUs. Not only that, but this GPU will enable LLMs (like GPT 4) to be trained up to 30 times faster.[33] This matters a great deal because computing power is one of the key factors in training AI models to be more apt at achieving the goals set out for them.[34] To capture this idea mathematically, here is a formula. This formula is not accurate, but it does a job of summing up the concepts and their interplay with AI Strength (strength being aptitude and intelligence). These are all very difficult to accomplish tasks and defining many of these terms is beyond me, but I want to use this formula to demonstrate an idea.
$S = C^{\alpha} * D^{\beta} * A^{\gamma}$

Here, α, β, and γ are positive constants that represent the relative importance of each factor (computational power, data, and algorithm efficiency, respectively). S is the strength of the AI system.

With enough compute, AI can achieve many many tasks, even with a bad algorithm and little data. As we've seen with brute force attacks in cybersecurity, brute force can accomplish many things.
Considering that even one of the older Large language AI models like GPT-3 have exhibited "emergent" capabilities already, who knows what we might see from newer language models such as those beyond GPT 3, 3.5, and GPT 4.[35]

Given these developments in technology, it is very unlikely that we will lack the capacity to provide a strong AI system with enough compute to be trained. It would seem that even if premise 1 from the Computational Limit argument is true, it is unlikely premise 2 holds and so the argument against misaligned AI from lack of computational power is shown to be likely unsound.

---

[32] A Graphics Processing Unit (GPU) is a specialized electronic chip designed to rapidly process and display images or videos, often used to enhance the performance of computers, gaming consoles, and other devices. In AI and large language models, GPUs are used to accelerate complex computations and parallel processing tasks, enabling faster training and more efficient execution of these models.

[33] "Nvidia Unleashes Its next-Generation Gpus, Dpus and Ai Accelerators," SiliconANGLE, March 21, 2023, https://siliconangle.com/2023/03/21/gtc-2023-nvidia-unleashes-next-gen-gpus-dpus-ai-accelerators/#:~:text=According%20to%20the%20company%2C%20the,faster%20AI%20inference%20on%20LLMs.
[34] "Compute, data, and algorithmic advances are the three fundamental factors that guide the progress of modern Machine Learning (ML)."
Jaime Sevilla et al., "Compute Trends across Three Eras of Machine Learning," arXiv.org, March 9, 2022, https://arxiv.org/abs/2202.05924.
[35] An ability is emergent if it is not present in smaller models but is present in larger models per, Jason Wei et al., "Emergent Abilities of Large Language Models," arXiv.org, October 26, 2022, https://arxiv.org/abs/2206.07682.

e. The Manual Specification Approach

You won't know until you study the influence of scale what capabilities or limitations might arise.

Deep Ganguli, Anthropic

The manual specification or Naive control approach to AI alignment assumes we can specify exactly what we'd like an AI system to do, and this system will always follow these instructions without any unintended or emergent consequences. It can be formally expressed as;

Argument For Manual Specification Approach (and counterargument):

(1) If we can specify exactly what we'd like an AI system to do (P), and this system will always follow these instructions without any unintended or emergent consequences (Q), then the manual specification approach is adequate for AI alignment (R).

(2) We can specify exactly what we'd like an AI system to do (P).

(3) This system will always follow these instructions without any unintended or emergent consequences (Q).

—----------

Therefore, the manual specification approach is adequate for AI alignment (R). (From 1, 2, and 3, Modus Ponens)

However, the premises (2 and 3) in the argument for the manual specification approach are problematic.

In LLMs such as GPT3, emergent capabilities have been shown to occur when trained with more parameters. Researchers generally have two explanations for these emergent capabilities in LLMs. The first is that larger models do in fact spontaneously acquire new skills as they are scaled up.[36] The second is that LLMS might be only capable of learning heuristics from more parameters and higher quality data.[37] Either way, emergence gives rise to unpredictable behaviors, which makes it difficult for researchers to anticipate the consequences of scaling up models. This implies that the manual specification approach isn't practical as LLMs continue to scale up in size due to the emergence of behaviors that weren't previously anticipated. "But what about different models? Surely AGI won't necessarily be from LLMs." This is true, it is not necessary that AGI will take the form of LLMs that have been scaled up. However, LLMs have jettisoned much of the current AI surge and have contributed significantly towards AGI by emulating human reasoning through the nature of their models.[38]

---

[36] Stephen Ornes and substantive Quanta Magazine moderates comments to facilitate an informed, "The Unpredictable Abilities Emerging from Large AI Models," Quanta Magazine, March 16, 2023, https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/.

[37] Ibid

[38] more on LLMs to come

Even with different models, the manual specification approach appears to be an improbable solution. While in theory it is possible to program a perfectly rational agent who can obey specific commands with perfect rationality, it is impractical to expect humanity to create a set of rules that are applicable in all scenarios. Even our own best moral theories lead to repugnant conclusions given the right settings for wrong outcomes.

Argument Against Manual Specification Approach:

(4) If an AI system can exhibit emergent behaviors, adaptability, and generalization that are hard to predict and control (P), then creating a comprehensive list of instructions for them is challenging due to inherent complexity and potential introduction of human biases (Q).

(5) AI systems, especially LLMs, do exhibit emergent behaviors, adaptability, and generalization that are hard to predict and control (P).

(6) Therefore, creating a comprehensive list of instructions for AI systems is challenging due to inherent complexity and potential introduction of human biases (Q). (From 4 and 5, Modus Ponens)

(7) If creating a comprehensive list of instructions for AI systems is challenging (Q), then the manual specification approach is inadequate for AI alignment (R).

—----------

Therefore, the manual specification approach is inadequate for AI alignment (R). (From 6 and 7, Modus Ponens)

C. Large Language Models and Superintelligence

a. Why would a text prediction model scale towards superintelligence?

i. Human Beings Compared to LLMs

Considering LLMS are essentially text prediction engines, it may seem like a leap to suggest they could become super intelligent agents that might spell doom for humanity. This question is valid and it is worth explaining why a text prediction model would scale towards superintelligence given enough compute, data, etc.

First, let us set aside the question of whether current architectures can scale to the required extents and focus on prediction models in general. In theory, sufficiently strong optimization on the task of text prediction has the potential to create intelligence on a scale far beyond human prediction and reasoning. This potential arises from several contributing factors.

LLMs are trained on enormous amounts of data, equivalent to far more text any human could read in a lifetime. GPT 3 was trained on 45 terabytes of raw text, and 570 gigabytes after the text was preprocessed.[39] For reference, the Library of Congress houses close to 51 million cataloged books.[40] If you were to take all of the text in the Library of Congress, you'd have around 10 terabytes of raw text, which is still four times less than the raw text used to train GPT 3. We still don't even know how much data was used to train GPT-4.[41] To contrast this, most calculations estimate that the average human reads around 700 books in their lifetime, so in some sense, GPT 3 has read orders of magnitude more books than you and I combined.

It is important to note that there are key differences between human beings and AI models that are overlooked when comparing sheer numbers of data processed by the systems. GPT 3 may have been trained on 45 terabytes of raw text, but this is all the data that model ever experiences, aside from the reinforcement it receives from human actors.

On the contrary, human beings are embedded in the environment and new experiences change our minds on a daily basis. In other words, human beings have a body, and multiple modes of perception. We are always taking in new experiences and "data" in some sense. Aside from the meager 700 books we read in a lifetime, estimates claim we absorb 34 gigabytes of data from experiences everyday.[42] It's difficult to say how much of this information is stored, but estimates place the capacity of the human brain somewhere around 1 - 2.5 petabytes or two million five hundred thousand gigabytes, the equivalent of three million hours of TV shows.[43]

This estimate should be taken somewhat lightly because the human brain is not digital, and so calculations based on digital estimates don't directly translate. We also don't know for sure how much of the daily information feed is saved and incorporated, and how much of it is "junk." This doesn't detract from the value of the point that's being made here, information is everywhere for a creature embedded in the environment, and one difference between data for an LLM and a human being is that the LLM is limited in the scope of the data it receives. If nothing else, hopefully this discussion has shown that it is very difficult drawing comparison between the human mind and LLMs, but there are insights that can arise even in the absence of a perfect comparison.

ii. Scaling up LLMs

While the last section argued that perhaps human beings are exposed to a wider breadth of data, it is unlikely this will be the case for long. As they are scaled up, and better LLMs are created, these models will be exposed to mind-boggling amounts more data than we currently are. In addition, they aren't subject to the same encoded rules that human beings have written

---

[39] Ahmed Mandour, "GPT-3.5 Model Architecture," OpenGenus IQ: Computing Expertise & Legacy (OpenGenus IQ: Computing Expertise & Legacy, February 1, 2023), https://iq.opengenus.org/gpt-3-5-model.

[40] "General Information : about the Library : Library of Congress," The Library of Congress, accessed May 2, 2023, https://www.loc.gov/about/general-information/.

[41] Matt Raymond, "How 'Big' Is the Library of Congress?: Timeless," The Library of Congress, February 11, 2009, https://blogs.loc.gov/loc/2009/02/how-big-is-the-library-of-congress.

[42] This number, 34 gigabytes, comes from a report in 2010 by the University of San Diego California.

[43] 1. Kamal Al-Malah, "The Human Brain: Search for Natural Intelligence," *International Journal of Educational Policy Research and Review* 8, no. 6 (2021), https://doi.org/10.15739/ijeprr.21.026.

within their DNA. This isn't to say LLMs are immune to bias because bias still exists in the data, but LLMs aren't hard coded with heuristics and biases as the human beings are. As LLMs are fed more data, they can acquire considerable amounts of knowledge across various domains and can observe patterns in the world that human beings aren't privy to. The LLMs may be able to see past certain mental blocks that human beings have due to the difference in the architecture making up the system. Similar to how a human being given infinite time can learn an infinite number of things (excluding mental storage limitations), it is possible with enough data, the correct architecture, and enough training, even an LLM can become superintelligent.[44]
Recall the definition for intelligence in Section 0;

> "An agent's level of intelligence is its level of effectiveness at achieving its goals. More intelligent systems are better at taking actions that will maximize their utility function. The actual goals do not determine intelligence."

To achieve goals, one must predict what actions an agent needs to take to reach these goals. Instrumental goals must be set on the way to achieving the final goal.
From here we can derive our definition for a superintelligent agent (S), as one that would be exceptionally effective at achieving its goals, surpassing any human capability.

The Argument of LLM Scaling Up (LSU):
Put in a formal argument, here is how we can argue that LLMs scale to superintelligence;

(1) As LLMs scale up, they can acquire extensive knowledge and pattern recognition capabilities, surpassing human exposure to data. (Basic)
   - ★ If LLMs scale up, then they can surpass human exposure to data in acquiring extensive knowledge and pattern recognition capabilities. (Basic, implicit)

(2) LLMs possess the ability to transfer learning and generalize across various tasks and domains, allowing them to effectively pursue diverse goals. (Basic)

(3) LLMs can iteratively learn from mistakes and improve their performance over time, enabling continuous growth in their effectiveness at achieving goals. (Basic)

(4) A superintelligent agent is one that is exceptionally effective at achieving its goals, which predominantly involve prediction. (Basic)

(5) Given enough data, it is theoretically possible to predict anything, as true randomness does not exist in the universe. (Basic)

—----------

C) Therefore, given enough data, compute, the capacity for extensive knowledge acquisition, advanced pattern recognition, and the absence of true randomness in the universe, a text

[44] We might not even need the "correct/optimal" architecture to create superintelligence. It is theoretically possible that with enough brute force, even a less than optimal model could become so precise in pattern recognition that it could become super intelligent.

prediction model (specifically, an LLM) has the potential to scale towards superintelligence, as defined by its exceptional effectiveness at prediction and achieving ends. (MP 1, 1*, 2, 3, 4, 5)

A key limitation

If the previous section is correct in the argumentation, we've shown that given enough data, compute, and a decent enough algorithm, LLMs can scale up to superintelligence. There is one glaring issue with this argument that some of you may have caught. How does something that requires prompting, as the GPT models do, become an agent? GPT X, or some other sufficiently advanced model might have the capacity for superintelligence, but how does this threaten humanity without agency? While the previous argument establishes the potential for LLMs to scale to superintelligence with respect to their pattern recognition, and potential for goal acquisition, they are limited by their lack of agency.

For LLMs to pose a threat to humanity, they need prompting, whether in a manner that ends like the cornfield conundrum, or by a bad actor, or agency of their own.[45]

Here is an informal argument that might explain how LLMs could pose an existential risk to humanity given agency.

1. LLMs have the potential to scale up to superintelligence in terms of knowledge acquisition, pattern recognition, and generalization. (from conclusion of LSU)

2. LLMs are currently passive, prompt-driven models, lacking inherent agency or goal-oriented behavior.

3. To become agents, LLMs must be endowed with goal-driven architectures, enabling them to autonomously pursue objectives

4. The development and integration of goal-driven architectures and decision-making systems with LLMs are technically possible, given advances in artificial intelligence, robotics, and related fields.

5. An autonomous, goal-driven LLM with superintelligence could pose risks to humanity, as its goals might misalign with human values, leading to unintended consequences or harmful actions.

Conclusion (C): While the current generation of LLMs does not inherently pose a threat to humanity due to their lack of agency, the potential for future LLMs to integrate goal-driven architectures and decision-making systems could lead to the emergence of autonomous, superintelligent agents. These agents, if not carefully aligned with human values and safety precautions, could pose risks to humanity.

[45] A bad actor is simply an agent that intends to do harm

While agency isn't a necessary component of threats from LLMs, it would certainly add to the danger of a superintelligent system.

D. Takeaway

Stories and examples such as those above aren't unlikely by any means, nor is this the fringe stance on AI safety. In fact, there is a survey, "Expert Survey on Progress in AI" that was run in 2016, 2019, and 2022 that found that experts believe there is a 10% probability of extinction from human failure to control AI.[46]
Experts such as Joe Carlsmith have written entire papers outlining the possibilities of not only risks from AI, but Existential Risks. X-RIsks are those such risks that would lead to the permanent disempowerment or extinction of the human species.

Carlsmith's specific words at the end of his paper are the following;
"As far as I can tell, there is a disturbingly high risk (I think: greater than 10%) that I live to see the human species permanently and involuntarily disempowered by AI systems we've lost control over. What we can and should do about this now is a further question. But the issue seems extremely serious."[47]

Having examined and argued against the various misconceptions regarding alignment, it is my hope that the reader has gained a deep understanding of where Carlsmith and the many experts in the field are coming from. In this section we've examined the many examples of misalignment and argued against the Benevolent Intelligence Argument, the AGI in a box argument, the Ultron Argument, the Computational Limits argument, and the Manual Specification approach. We've detailed how language prediction models can scale up to superintelligence, and explored how, even without agency, these models can wreak havoc.

The section that follows will explain why, even after all the danger, many still seek to create AGI due to the potential that the field holds.
After reading the misalignment section, I hope the reader feels called to action. This is our call to investigate, to reach into the various fields and ideas, to examine them, and at the very least, to give them our attention in the pages that follow.

---

[46] "2022 Expert Survey on Progress in AI," AI Impacts, September 27, 2022, https://web.archive.org/web/20221016004611/https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.
[47] "Existential Risk from Power-Seeking AI (Shorter Version)," Joe Carlsmith, accessed April 20, 2023, https://joecarlsmith.com/2023/03/22/existential-risk-from-power-seeking-ai-shorter-version.

III. "Why AI?" Delving into the Drive for Strong AI Development

Curiosity is the essence of human existence. 'Who are we? Where are we? Where do we come from? Where are we going?'... I don't know. I don't have any answers to those questions. I don't know what's over there around the corner. But I want to find out.

Gene Cernan

This section explores the reasons why we might wish to pursue strong AI. It covers the general reasons such as financial incentives, aiding in research, advancing technology, curing diseases, solving global problems, and enabling space exploration. I also examine the relationship with human curiosity, hedonic adaptation, boredom, and the economy. It is my hope that by taking a step back to recognize the drivers behind AI, we might find more clarity in our behaviors.

A. Reasons for Pursuing Strong AI

a. The General Goods

The pursuit of AGI/strong AI, has several motivations. While there are a plethora of risks associated with its continued development, we must understand why many companies are steadfast in their endeavors. Here are a few of the key reasons for the pursuit of AGI;

1. Economic Benefits: History has shown us that the first group to a new technology often dominates all others. The newfound capital could fund further research and create a feedback loop leading to a near monopoly. If nothing else, the first company to develop AGI would reap immense financial rewards and gain a leg up over its competitors.
2. Geopolitical Advantages: The country that first creates AGI would be far more powerful than all others on the global front. Whoever holds the power often controls the direction of the future.
3. GDP, Labor, and Productivity: AGI could perform services across broad domains, freeing up leisure time, and allowing for things such as universal basic income.
4. Acting as our Oracle: AGI may serve as an oracle for humanity, answering complex equations, addressing issues in a variety of fields, and even leading to the development of new technologies and theories.
5. Enhancing Humanity: It is possible that AGI could be connected with human beings and allow us to augment/enhance our abilities.
6. Solving Global Issues: AGI could be the key technology that enables us to cure human diseases, fix supply chains, solve world hunger, and fix a variety of the issues that are faced on a global scale. With a technology that is more intelligent and faster than we are,

there are many possibilities for how it could be used to solve complex problems that human beings have been faced with.

7. Space Exploration: AGI could help us colonize the stars, thus spreading the light of consciousness throughout the universe.

Many of these reasons are based on the assumptions that AGI would be aligned, and sufficiently intelligent to solve issues that some of the most brilliant human beings have failed to solve, whether due to lack of knowledge or lack of time. It is possible that if we are able to develop truly aligned ASI/AGI, many of these points will take place. Only time will tell, but whether these reasons come true or not, the future of humanity hangs in the balance.

b. Ensuring Humanity's long term future

Aside from the many claims of what strong AI might bring for humanity, there is one more important reason for continuing its development that I left out.

8. Ensuring the Long-term survival of humanity by mitigation of existential risks.

If we can survive until the end of the universe, most of the human beings who will ever exist, are still yet to be born. This means that the future is one of the most important things because all that which any person loves will exist in more abundance in the future, then any other time in history. This claim leads us to a moral theory intertwined with the discussion of existential risks and humanity's future, long termism. This philosophy is outlined by William MacAskill, Oxford philosopher, in his book, *What We Owe the Future*. Long Termism is defined as "the view that positively influencing the long term future is a key moral priority of our time."[48] As I mentioned at the start of this subsection, the reasoning behind this claim is that most of the human beings who will ever live are to come after us, and we owe it to them to set up a future where their quality of life is good. The specifics of this philosophy are beyond the scope of this paper, but the point has been made and does pertain to this discussion. One need not agree with the long termist principles to concede the point that the claim carries truth.

In the long termist community, there are those who claim that this century is the most important in human history because it is the most likely we have ever faced where we are liable to be wiped out, this killing or disempowering all of us, and ruining the chances of a good future for all those that will come after. This claim is backed up by research done by Toby Ord. His estimate of the risk of possible extinction or permanent disempowerment in this century is 1 in 6 (17%).[49] These numbers represent the case in which things go along with the status quo. We are able to change them if we intervene now. That is the key point.

The numbers for the specific existential risks we face are as follows;

Existential catastrophe via	Chance within next 100 years
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000

[48] MacAskill, William (2019-07-25). "Long Termism". Effective Altruism Forum.

[49] The shocking thing is that if these numbers are true, humanity is essentially playing russian roulette. TOBY ORD, *Precipice: Existential Risk and the Future of Humanity* (S.l.: BLOOMSBURY PUBLISHING, 2020).

Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	~ 1 in 10,000
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
'Naturally' arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	~ 1 in 6
Total existential risk	~ 1 in 6

[50]

It is possible that with an aligned AGI/ASI we may stand a better chance of survival as this strong AI could help to mitigate various existential risks. I'll describe some of the ways a strong AI system might help humanity's long term survival, but to quantify the exact probabilities is beyond me.

1. Asteroid Impact: Chart potential asteroids using prediction powers far-surpassing the current abilities of human beings working with computers, develop early-warning systems, perform difficult physics and math calculations to create countermeasures
2. Supervolcanic eruptions: Improve predictions of eruptions, generate novel solutions for preventing the eruptions (if possible)
3. Nuclear war: utilize knowledge of human nature, model predictions for diplomatic relations, aid in diplomacy, help manage nuclear arsenals, act as a near perfectly rational agent to do the greatest possible good
4. Climate Change: This field faces issues with renewable energy, supply chain, carbon capture, and the extreme complexity of modeling climate due to the sheer number of variables involved. Strong AI, with fast computation, could potentially solve these issues.
5. Bio-risks: AGI could potential be used to aid in the development of tools for controlling outbreaks, generation of novel treatments, predicting the next pandemic, and even doing rapid gain of function predictions to anticipate what viruses might evolve and how we can eliminate them
6. Unforeseen anthropogenic risks: "Anything that Can Go Wrong, Will Go Wrong."[51]Murphy's law, while not a law of nature, helps to illustrate the point that there will be consequences we don't intend for and cannot anticipate. Perhaps AGI/ASI would be capable of predicting X-risks caused by humanity, and aiding us in mitigating them long before they occur.

---

[50] Taken from ibid

[51] Author(s) D D Garwood, "Edsel Murphy's Law - Anything That Can Go Wrong, Will Go Wrong - Murphy's General Law as Applied to Program Evaluation," Edsel Murphy's Law - Anything That Can Go Wrong, Will Go Wrong - Murphy's General Law as Applied to Program Evaluation | Office of Justice Programs, accessed May 4, 2023, https://www.ojp.gov/ncjrs/virtual-library/abstracts/edsel-murphys-law-anything-can-go-wrong-will-go-wrong-murphys.

If strong AI could be aligned, it may very well mean that much of these issues are solved and the payoffs could be astronomical. Strong AI could usher in the era of prosperity and exploration of the cosmos, an era the likes of which humanity has never seen before. As amazing as this period may be, this doesn't necessarily mean the end of human suffering. There is something inherent to humanity that may prevent us from ever-achieving that end utopia that so many have dreamed of.

B. AI as a Manifestation of Human Curiosity and Problem-Solving Drive

The development of strong AI can be viewed as anecdotal evidence for the human quest for knowledge and progress-evermore. As alluded to in the previous section, there is something that may hinder the human quest for happiness, even with aligned superintelligence giving us all that we wish for. In this section, I argue that our pursuit of strong AI is deeply rooted in our evolutionary history, and as such, it won't disappear even after we've solved world hunger, conquered the stars, and created a utopia.
I call this informal argument, *The Hedonic Treadmill Argument (**HTA**)*;

(1) Human beings have an innate curiosity and drive to solve problems, hard-coded into their DNA that will not change over time.
(2) Human beings adjust to new circumstances and return to a baseline level of satisfaction, continually seeking improvements and new experiences.[52] This is known as hedonic adaptation.
(3) AGI will not change human DNA, but will help to make the external circumstances and living conditions better and better for human beings.

—----------

Conclusion: Given that human beings possess an innate curiosity and problem-solving drive, along with the tendency towards hedonic adaptation, unless something alters this, we will still possess these traits even after a beneficial societal shift from aligned strong AI. Strong AI may significantly improve our external circumstances and living conditions, our inherent drive for knowledge, exploration, and new experiences will persist, continually pushing us to seek further advancements and discoveries.

b. Curiosity and The Human Condition

To illustrate the impact that curiosity has in this world, I will lay out some statistics.

[52] In the field of the psychology of happiness, much research has been done on this phenomenon. Evidence has shown that regardless of most life events, the subjective happiness of people returns to baseline after a long enough period.
"Hedonic Adaptation," Hedonic Adaptation - an overview | ScienceDirect Topics, accessed May 4, 2023, https://www.sciencedirect.com/topics/psychology/hedonic-adaptation.

Currently, there are 3.5 billion web searches per day.[53] Of these searches, around 8% are in the form of a question.[54] This means there are nearly 280 million questions asked on the internet each day. The general dictionary consensus on what constitutes a question is a request for information/knowledge.[55] There is no shortage of human beings asking questions. It's safe to say that curiosity is a part of the human experience, but is it a key driving force behind our desire to create AI?

Aside from the time spent directly in activities necessary for our biological survival, much of our time "is spent seeking and consuming information, whether listening to the news or music, browsing the internet, reading books or magazines, watching TV, movies, and sports, etc."[56] Why do we spend so much time seeking information? William James thought this curiosity was "the impulse towards better cognition" roughly translating to the desire to understand what you don't know.[57] One study on rats revealed they engage in non-specific curiosity in the absence of specific tasks such as searching for food, looking for a mate, etc. The rats began exploring areas of the maze that they were unfamiliar with.[58] How does this connect to AI and human beings? In some sense, we're like the rats, searching and exploring for its own sake. Our journey to learn and do more is emergent from the innate curiosity honed over millions of years. This isn't to say that the creation of AI doesn't pose possible benefits for those who are in need of the essentials for life, rather I claim that the individuals who are curious about AI cannot help but be that way due to the inherent nature of curiosity. It is so innately human to push the boundaries in some way, and to seek information in a curious manner. Information seeking doesn't necessarily mean reading long-form text, scholarly content, or even anything that seems informative. Many activities are sufficient to qualify for curiosity and information seeking. Watching TV, reading romance novels, going on trips, any novelty seeking behavior is, for the most part, curiosity.

C. Takeaway

In this section we've covered the primary motivations human beings would want to pursue strong AI. We've looked at the general benefits strong AI can provide, the financial incentives, and how strong AI might enable the flourishing of humanity long into the future. We've also examined how AI is the product of innate human characteristics such as the drive to solve-problems, and curiosity. This section was important because to gain an intuition about the AI space, we need to know why we're on this journey to begin with.

---

[53] "Google Search Statistics 2023," TrueList, January 9, 2023, https://truelist.co/blog/google-search-statistics/.
[54] Si Quan Ong, "78 Seo Statistics for 2023," SEO Blog by Ahrefs, November 15, 2022, https://ahrefs.com/blog/seo-statistics/.
[55] "Question Definition & Meaning," Merriam-Webster (Merriam-Webster), accessed May 4, 2023, https://www.merriam-webster.com/dictionary/question.
[56] Celeste Kidd and Benjamin Y Hayden, "The Psychology and Neuroscience of Curiosity," Neuron (U.S. National Library of Medicine, November 4, 2015), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4635443/.
[57] Ibid
[58] (e.g., Dember, 1956; Hughes, 1968; Kivy, Earl, & Walker, 1956)

IV. Conclusion

A. Recap of Key Arguments and Perspectives

This paper has presented both faces of the AI space. Neither topic was explored in the full-depth, and it was never my intention to do so. Rather this paper was meant to serve as a guide for building intuitions about AI, and a map for understanding both sides of a debate that is making headlines throughout the globe. We've explained what intelligence is, explored the alignment problem, set arguments up, and knocked them down. On the contrary, we've delved into reasons for pursuing strong AI, and why this pursuit is occurring at a deeper level, namely, that of psychology and DNA.

B. The Importance of Now

In his book, *What We Owe the Future,* William Mackaskill highlights a concept known as "early plasticity, early rigidity." This refers to periods of great change, where the world is highly plastic at times of turmoil and shifting dynamics. As things settle down, ideas and beliefs become more and more locked in and things are difficult to change. This period of humanity finds itself in such a shifting stage. The DotCom boom was less than a quarter century ago, and the current AI boom is the beginnings of large changes to come. As we move forward, it's critical that we don't lock ourselves into a bad scenario. As Mackaskill puts it, ""what's at stake when navigating the transition to a world with advanced AI, then, is not *whether* civilization continues, but *which* civilization continues."[59] This begs the question, how do we avoid locking ourselves in a bad world? We need to acknowledge that we aren't right about everything, and if the past proves anything, it is that we've been guilty of thinking our moral and social theories were perfect, even when they'd be considered egregiously wrong today.[60] We must be well educated and question our belief systems, even if we don't want to. If virtues do exist, maybe willingness to question, or openness to experience ought to be one.

As this paper illustrates, it's important to be aware of the many faces of an argument so we can understand where we came from, and from there, plot the course of where we want to go. That is where the title comes from, "Facing Janus." In Roman lore, Janus is the god of beginnings, endings, and transitions, often depicted with two faces that look in opposite directions. We need to be looking to our past for examples, and looking forward to where we

[59] William MacAskill, *What We Owe the Future* (Basic Books, 2022).
[60] Think slavery, beliefs in the Geocentric universe, Newton's laws as a perfect description of the universe… we've been wrong so many times before,it is unlikely we've got it perfect this time.

want to go, but with our primary attention rooted in the decisions we have available to make now. This is a difficult feat, and one that will take considerable effort. With that being said, the next subsection illustrates potential ideas for ensuring a safe future.

C. Final Recommendations for AI Policy and Development

It was my original intent to go through mathematical calculations to establish how "good" the world would be given different scenarios such as; AI enslavement, AI alignment success, mitigating X-risks without AI, etc… but eventually I decided against this as too many variables are left unknown to me. Rather than add another set of expected value calculations to a field teeming with them, I think it would be a more fitting end to this paper to lay out a number of broad recommendations and topics that should be examined more closely by myself or someone else in further research.

Specific Recommendations:

1. Focus on the development of Artificial Specific Intelligence tools (ASI)
    a. Calculate likelihood of X-Risk Mitigation with ASI, and without AI
    b. make ASI a priority that provides domain specific solutions without the risks
2. Enhance education on psychology and cognitive biases
3. Halt all AGI mass testing until humanity has the necessary ability to align LLMs or whatever te main model is at that time[61]
4. Scale up studies of the human mind, biases, the Spinozan model of the mind:
    a. Increase funding and support for research into the Spinzoan model of the mind and its effects on human behavior.[62]

General Recommendations:

1. **Public Awareness on a deeper level:** Much of this paper was spent engaging with arguments against misalignment. This was because from my own experience many, even those well-educated in computer science, simply don't understand how we could be harmed by AI. The intuition just isn't there. We need to educate the public on alignment issues, and every little bit helps. 80,000 hours and others in the effective altruist space have done well with this. We need continuous discourse and public voices spreading the message of the dangers of AI.
2. **Prevent value lock-in:** Value lock in, as touched on earlier, is when an idea is "locked" in for society and it stifles the ability for deviation from one belief. Examples of this are the Spanish Inquisition, and the North Korean Regime currently forcing propaganda and

[61] Bold claim, but I hope the reader can see where I am coming from after this paper

[62] In the paper "Thinking is Believing" by Eric Mandelbaum, he presents and supports the Spinozan model of the mind, which posits that the human mind automatically accepts propositions as true. This work and the concept of this model have significant implications for understanding human cognition and belief formation. This can aid in understanding epistemology and hopefully what human beings value, based on how their beliefs and values form.
Eric Mandelbaum, "Thinking Is Believing," Inquiry 57, no. 1 (2013): pp. 55-96, https://doi.org/10.1080/0020174x.2014.858417.

limiting information that goes against their current narrative. “When we look at history, we see that the predominant culture in a society tends to entrench itself, eliminate competition, and takes steps to replicate itself over time.”[63]

3. **Be Open-minded, yet discerning:** In chaotic times, it is tempting to seek solace in theories providing clean cut, black and white explanations about the world. However, we must guard our minds against this, because reality is not simple, and certainly not black or white. A good rule of thumb as Peter Godfrey-Smith puts it, “Beware the dubious allure of simplicity in philosophical theories,” strive to maintain critical-thinking and intellectual humility.[64]

D. Closing Statements

While the recommendations I presented may seem too broad or perhaps even obvious, their importance shouldn’t be understated. As the tech industry continues its pursuit of strong AI, alignment research needs to scale much faster than it currently is. The public needs to be educated, and philosophical intuitions need to be acquired by those in power.
The pursuit of strong AI carries the potential to bring about the best or the worst time in human history, and even if it doesn’t go to extremes, it will certainly bring a dramatic shift to the status quo. By focusing on education, research, and open and honest discourse, it is possible humanity can navigate this harrowing situation and come out on top. For this to take place, we MUST find a way to align AI in such a way that under nearly no circumstances can dangerous misalignment occur. This is incredibly difficult, but we can get closer to this goal through rehashing ideas, building up arguments, and tearing them down. This will take a collaboration of scientists, philosophers, politicians, educators, and of course, the public.

“What we do in life, echoes in eternity.”
- The Stoic Emperor,
Marcus Aurelius

\

[63] William MacAskill, *What We Owe the Future* (Basic Books, 2022).
[64] Peter Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science* (Chicago (Ill.): The University of Chicago Press, 2021).

Bibliography

"2016 Expert Survey on Progress in AI." AI Impacts, March 26, 2021. https://aiimpacts.org/2016-expert-survey-on-progress-in-ai/.

"2022 Expert Survey on Progress in AI." AI Impacts, September 27, 2022. https://web.archive.org/web/20221016004611/https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Al-Malah, Kamal. "The Human Brain: Search for Natural Intelligence." *International Journal of Educational Policy Research and Review* 8, no. 6 (2021). https://doi.org/10.15739/ijeprr.21.026.

Author(s) D D Garwood. "Edsel Murphy's Law - Anything That Can Go Wrong, Will Go Wrong - Murphy's General Law as Applied to Program Evaluation." Edsel Murphy's Law - Anything That Can Go Wrong, Will Go Wrong - Murphy's General Law as Applied to Program Evaluation | Office of Justice Programs. Accessed May 4, 2023. https://www.ojp.gov/ncjrs/virtual-library/abstracts/edsel-murphys-law-anything-can-go-wrong-will-go-wrong-murphys.

"Dictionary by Merriam-Webster: America's Most-Trusted Online Dictionary." Merriam-Webster. Merriam-Webster. Accessed April 19, 2023. https://www.merriam-webster.com/.

"Existential Risk from Power-Seeking AI (Shorter Version)." Joe Carlsmith. Accessed April 20, 2023. https://joecarlsmith.com/2023/03/22/existential-risk-from-power-seeking-ai-shorter-version.

"General Information ： about the Library ： Library of Congress." The Library of Congress. Accessed May 2, 2023. https://www.loc.gov/about/general-information/.

Godfrey-Smith, Peter. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago (Ill.): The University of Chicago Press, 2021.

"Google Search Statistics 2023." TrueList, January 9, 2023. https://truelist.co/blog/google-search-statistics/.

Haan, Hein de. "The AI Box Experiment." Medium. The Singularity, October 1, 2020. https://medium.com/the-singularity/the-ai-box-experiment-c92a0a389eb7.

"Hedonic Adaptation." Hedonic Adaptation - an overview | ScienceDirect Topics. Accessed May 4, 2023. https://www.sciencedirect.com/topics/psychology/hedonic-adaptation.

Hilton, Benjamin. "Preventing an AI-Related Catastrophe - Problem Profile." 80,000 Hours, March 28, 2023. https://80000hours.org/problem-profiles/artificial-intelligence/.

"How AI and New Technologies Reinforce Systemic Racism - Ohchr.org." Accessed April 21, 2023. https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/advisorycommittee/study-advancement-racial-justice/2022-10-26/HRC-Adv-comm-Racial-Justice-zalnieriute-cutts.pdf.

Kidd, Celeste, and Benjamin Y Hayden. "The Psychology and Neuroscience of Curiosity." Neuron. U.S. National Library of Medicine, November 4, 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4635443/.

Lacey, Heather P, Angela Fagerlin, George Loewenstein, Dylan M Smith, Jason Riis, and Peter A Ubel. "Are They Really That Happy? Exploring Scale Recalibration in Estimates of Well-Being." Health psychology : official journal of the Division of Health Psychology, American Psychological Association. U.S. National Library of Medicine, November 2008. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3744864/.

MacAskill, William. *What We Owe the Future*. Basic Books, 2022.

Mandelbaum, Eric. "Thinking Is Believing." *Inquiry* 57, no. 1 (2013): 55–96. https://doi.org/10.1080/0020174x.2014.858417.

Mandour, Ahmed. "GPT-3.5 Model Architecture." OpenGenus IQ: Computing Expertise & Legacy. OpenGenus IQ: Computing Expertise & Legacy, February 1, 2023. https://iq.opengenus.org/gpt-3-5-model.

Martin, Sammy. "Investigating AI Takeover Scenarios." LessWrong. Accessed April 20, 2023. https://www.lesswrong.com/posts/zkF9PNSyDKusoyLkP/investigating-ai-takeover-scenarios#Table.

"A New AI Lie Detector Can Reveal Its 'Inner Thoughts.'" Big Think, April 3, 2023. https://bigthink.com/the-future/ai-lie-detector/?utm_campaign=later-linkinbio-bigthinkers&utm_content=later-34606118&utm_medium=social&utm_source=linkin.bio.

"Nvidia Unleashes Its next-Generation Gpus, Dpus and Ai Accelerators." SiliconANGLE, March 21, 2023. https://siliconangle.com/2023/03/21/gtc-2023-nvidia-unleashes-next-gen-gpus-dpus-ai-accelerators/#:~:text=According%20to%20the%20company%2C%20the,faster%20AI%20inference%20on%20LLMs.

Ong, Si Quan. "78 Seo Statistics for 2023." SEO Blog by Ahrefs, November 15, 2022. https://ahrefs.com/blog/seo-statistics/.

Ong, Si Quan. "78 Seo Statistics for 2023." SEO Blog by Ahrefs, November 15, 2022. https://ahrefs.com/blog/seo-statistics/.

ORD, TOBY. *Precipice: Existential Risk and the Future of Humanity*. S.l.: BLOOMSBURY PUBLISHING, 2020.

Ornes, Stephen, and substantive Quanta Magazine moderates comments to facilitate an informed. “The Unpredictable Abilities Emerging from Large AI Models.” Quanta Magazine, March 16, 2023. https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/.

“Question Definition & Meaning.” Merriam-Webster. Merriam-Webster. Accessed May 4, 2023. https://www.merriam-webster.com/dictionary/question.

Raymond, Matt. “How 'Big' Is the Library of Congress?: Timeless.” The Library of Congress, February 11, 2009. https://blogs.loc.gov/loc/2009/02/how-big-is-the-library-of-congress.

Reber, Paul. “What Is the Memory Capacity of the Human Brain?” Scientific American. Scientific American, May 1, 2010. https://www.scientificamerican.com/article/what-is-the-memory-capacity/.

Russell, Stuart J., Peter Norvig, and Ernest Davis. *Artificial Intelligence: A Modern Approach*. Harlow, England: Pearson Educación, 2022.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. “Compute Trends across Three Eras of Machine Learning.” arXiv.org, March 9, 2022. https://arxiv.org/abs/2202.05924.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. “Emergent Abilities of Large Language Models.” arXiv.org, October 26, 2022. https://arxiv.org/abs/2206.07682.

“Worldwide Spending on AI-Centric Systems Will Pass $300 Billion by 2026, According to IDC.” IDC. Accessed April 20, 2023. https://www.idc.com/getdoc.jsp?containerId=prUS49670322.

Yudkowsky, : Eliezer S. “The AI-Box Experiment:” Eliezer S Yudkowsky, July 2, 2021. https://www.yudkowsky.net/singularity/aibox.