AGI Analysis

Peter Parker - May 2023

1. Introduction

This paper details my thoughts and reasoning on the intersection of intelligence with biological and artificial life. As with all sciences, no single analysis or opinion can speak for an entire field. Nonetheless, I believe it is beneficial to express my ideas and opinions, even if some components may be debatable or even incorrect.

This paper aims to guide readers in their thought process on the topic of AGI and adjust their opinions in logical ways to gain a stronger, new perspective. The use of phrases like "we think" indicates my best interpretation of facts and ideas from experts in the field, which can generally be considered true. Conversely, the use of phrases like "I personally believe" pertains solely to my intuition and perspective on the subject.

Throughout this paper, I use my own definitions of certain words. These definitions may not always align with those found in Webster's dictionary or the scientific consensus. However, I use them in a manner that I find advantageous within the context of this paper.

It is important to note that throughout this paper, I frequently refer to animals. When using the term "animals," I am referring specifically to non-human animals, even though humans are also classified as animals.

The topic of AGI is something that I find extremely intriguing and immensely significant. It has motivated me to create this paper of ideas, and while I may have limited experience with AI and AGI, I believe that the propositions and logical arguments presented here are reasonable.

Despite humans having ruled the Earth for thousands of years, it is important to know that all humans are stupid. Obviously there are countless arguments against this information, but in the grand scheme of the universe and the potential for intelligence it is true. Our capacity to accumulate data is finite, our memories are significantly restricted, our understanding of complex models is limited, we are susceptible to biases, and we make mistakes. In contrast, AGI can reach extraordinary levels of intelligence. All humans are stupid.

2. Questions

This paper will address the following questions. By reviewing them, readers can familiarize themselves with the topics to be discussed and obtain a general understanding of the paper's content:

3. What is AGI?
4. What makes something important?
5. Are computers intelligent?
6. Are animals intelligent?
7. Do computers have general intelligence?
8. Do animals have general intelligence?
9. Why are humans more intelligent than animals?
10. What does it mean to be living?
11. What does it mean to be sentient?
12. Why is coding general intelligence hard?
13. What is math?
14. What is problem solving?
15. What is science?
16. What are the dangers of creating AGI?
17. What are the dangers of not creating AGI?
18. What does it mean to understand?
19. What does it mean to understand language?
20. Why is communication with humans easier than with computers and animals?
21. What is a pattern?
22. How do we generalize examples?
23. What is engineering?
24. When will we create AGI?
25. When will we know we have created AGI?
26. What is creativity?
27. Who will create AGI?
28. How fast will AGI take off?
29. What is curiosity?
30. What happens when AGI becomes more intelligent than humanity?
31. What should future life with AGI look like?

3. What is AGI?

AGI, or artificial general intelligence, refers to a general intelligent machine that is not derived from biological entities such as humans or animals, but instead created using computers or other technologies. **Intelligence** is the ability to perform important tasks with accuracy and efficiency. Similarly, **learning** is the improvement of intelligence in a task. **General intelligence** is the ability to perform a wide range of important tasks with accuracy and efficiency.

AGI is often described as the algorithm or program that will surpass human intelligence and accelerate our technological advancement towards utopia, dystopia, or extinction. The belief is that computers can process and calculate information at a much faster rate than humans, and can improve upon many of our inefficiencies. Some examples of these human inefficiencies include slower electrical speed of neurons, limitations in intelligent thinking during the early years of life, the time used for restorative sleep, the necessity for individuals to relearn knowledge already acquired by humanity, and the potential for mistakes due to inattention or natural imperfect behavior. It is evident that AGI has the potential to significantly enhance our intellectual capabilities compared to humans.

In addition, **friendly AGI** is an AGI that is designed to benefit civilization, leading us towards a utopian future rather than a dystopian one or extinction.

4. What makes something important?

An idea is **important** if it is repeatedly encountered, has the perception of a shift in world perspective, or is explicitly stated as important by a credible source.

Repetition indicates relevance, as ideas that are frequently encountered are more likely to continue to arise. For instance, knowledge about the color of the sky is more important than knowing the square root of 12,938,409 is 3,597 since the latter information will rarely ever come up.

The importance of something can also be relative. For example, if you were to give a smartphone to a primitive tribe, they may find it novel but ultimately useless and unimportant. However, if you were to give that same smartphone to a young American teenager, they would likely be ecstatic for the free gift as it is important to them. The gift of a smartphone can change a teenager's perception on life significantly because it provides them with access to

communication with friends and family, entertainment, internet information, and much more. Therefore, importance can vary depending on the people and time within context.

At times, something may seem unimportant, but in actuality it is important. This is especially true when an expert or an individual with extensive knowledge on a subject asserts its importance. Consider the example of a parent telling their child to eat their veggies, which the child may not view as important since they would rather eat cookies. However, eating veggies leads to a healthier lifestyle and better life outcomes. The complexities of this may be difficult to explain to a child, but it doesn't diminish the importance of the advice.

5. Are computers intelligent?

Computers are designed to perform algorithms that produce accurate actions and optimizations, provided they are programmed correctly and with good **dummy proofing** (the process that handles unexpected inputs to prevent incorrect behavior). However, since computer programs are created by humans, mistakes are possible, and these errors can propagate through multiple systems since programs typically use other programs as components. To minimize mistakes, programmers use systematic testing and user feedback to target mistakes and prevent incorrect outputs. Some algorithms are mathematically built from the ground up, resulting in provable correct programs. However, in the real world there can be human and hardware mistakes that make certainty imperfect. Despite all of these possible imperfections, in theory computers can perform most tasks optimally when following the correct algorithm. In practice, computers have been shown to outperform humans across many domains of tasks. Although computers are not superior to humans in every task, computers are more intelligent than humans across many domains.

6. Are animals intelligent?

Yes, but to a significantly lesser degree than humans. But why is this the case? There are numerous theories, including differences in brain size and brain-to-body ratio, disparities in the prefrontal cortex, variations in language capabilities and tool-making abilities, and the plasticity of the brain, among other ideas.

I believe that our superior communication skills and the plasticity of our brains are the key factors that contribute to our dominance in intelligence. These abilities allow us to learn

collaboratively, abstract complex ideas, and adapt our models of the world throughout our lifetime.

However, one aspect that is often overlooked is the intellectual capacity of animals. Many people assume that certain abilities are exclusive to humans, but in reality, they also exist in the animal kingdom. Both humans and non-human animals can create tools, communicate, problem-solve, deceive, anticipate the future, produce art, learn, recognize oneself, plan, and so on. While humans may excel in these areas, it is important to recognize the remarkable abilities of animals as well. Therefore, animals are intelligent across a wide range of similar intelligent tasks as compared with humans. More discussion on this topic will come later.

7. Do computers have general intelligence?

No. Currently, computers rely on human-created algorithms to intelligently solve problems. Computers cannot function or solve problems without human intervention. Therefore, when a computer solves a problem, it is a result of a human or a team of humans constructing an algorithm to solve that particular problem.

Currently, popular programs like GPT and Dalle-2 have shown remarkable intelligence in areas such as language and art, surpassing what was thought possible. These programs rely on massive amounts of data and deep neural networks, which enable them to generate outputs that rival human creations. Although these systems are still limited and not yet as accurate as human professionals, they are highly efficient. However, they require extensive amounts of data and computation which means these AI systems typically can only be created by large teams of people. Moreover, their application is restricted to problems that have enormous amounts of data, but this can be readily available today due to the internet's vast scale of information. Nevertheless, there are issues with these systems, including their black-box nature, which makes it challenging to debug the source of inaccurate outputs. Moreover, they lack the ability to use math, reason, or logic (which is ironic given that we perceive computers as the pinnacle of math, reason, and logic). Instead, these systems approximate logical thinking in ways that humans can’t always comprehend, and their solutions cannot always be correct. Most optimal algorithms exhibit a variable amount of time to compute an answer, but most neural network systems produce an output in a fixed time through a set number of matrix multiplications, indicating that they cannot possibly solve all algorithmic problems accurately. Despite these limitations, deep neural networks are still an invaluable tool and represent a significant step towards achieving AGI. For instance, ChatGPT is the editor used to review this paper.

Ultimately, the ability to maintain a job is a good measure of general intelligence, as it requires the capacity to understand and problem solve across various contexts. While any generic human can achieve any intellectual job, there is no universal algorithm that can replace all intellectual jobs. Therefore, we still don't have an AGI.

8. Do animals have general intelligence?

Yes. Animals can navigate through different terrains, forage for food, take care of their young, identify potential threats, communicate with each other, and adapt to harsh environments. Unlike computers, which rely on human intervention to learn, animals can figure out how to survive through their lifetime with only a small amount of data from interaction with their parents and social groups. Their ability to behave intelligently across various activities and environments is a testament to their general intelligence.

9. Why are humans more intelligent than animals?

Everything in this section reflects my personal beliefs regarding the answer to this question. Humans have built cars, planes, and cities. Humans have stepped foot on the moon and sent robots to some of the glowing dots in the night. Humans have developed vaccines, created the internet, and made nuclear bombs. Animals ain’t done shit. Don’t get me wrong, I love animals and they are very important and should be valued, but they are not as intelligent as humans.

I believe that the reason humans are more intelligent than animals lies in our superior communication skills and the plasticity of our brains. Humans have developed hundreds of spoken, written, and signed languages that together comprise millions of names, words, and ideas. Each of these words can be combined to form exponentially more sentences, phrases, and complex ideas. In this paper, I am only using a fraction of these words, yet I can still convey complex ideas to you. While animals have some communication capabilities, no animal comes close to matching our language abilities. Communication is essential for intelligence because it allows knowledge, skills, and tools to be shared among humans. This facilitates the spread of ideas from both living and deceased humans, leading to the exponential growth of knowledge that we can apply to many aspects of our world.

One notable aspect to consider is that communication between humans and animals is much more limited than human-to-human communication. A compelling case study is Koko the

gorilla, who was able to use sign language to attain communication skills that rivaled those of a developed adult. However, it took a dedicated researcher many years to accomplish this, whereas children learn communication skills more naturally and with less attention.

Similarly, communication between humans and computers can be challenging and time-consuming, particularly when developing interfaces to communicate. The more complex the interaction, the more time-consuming to develop the interface to account for the wide range of user actions. Moreover, computers are more vulnerable to input errors, which can lead to unintended consequences. Even a single character change in the input can cause significant output errors. To prevent this, computer programmers must spend large amounts of time dummy proofing the input, and this time detracts efficiency within human to computer communication.

Additionally, the plasticity of the brain is a crucial factor in how we learn. The saying "you can't teach an old dog (or human) new tricks" stems from the fact that as humans age, the plasticity of the brain decreases (but it’s not entirely lost, so individuals can still acquire new knowledge and skills even in their later years). Children, on the other hand, have a much higher level of brain plasticity than adults, which enables them to learn new things at a significantly faster rate. Humans, in general, exhibit a higher degree of brain plasticity than other animals, which could partially account for our higher intelligence. This ability to learn is essential for disciplines such as logic, mathematics, engineering, and science, which have facilitated humanity's impressive accomplishments.

Additionally, I believe intrinsic curiosity is another important component of human intelligence. This trait is essential for discovering new information and tools that expand human knowledge and abilities.

10. What does it mean to be living?

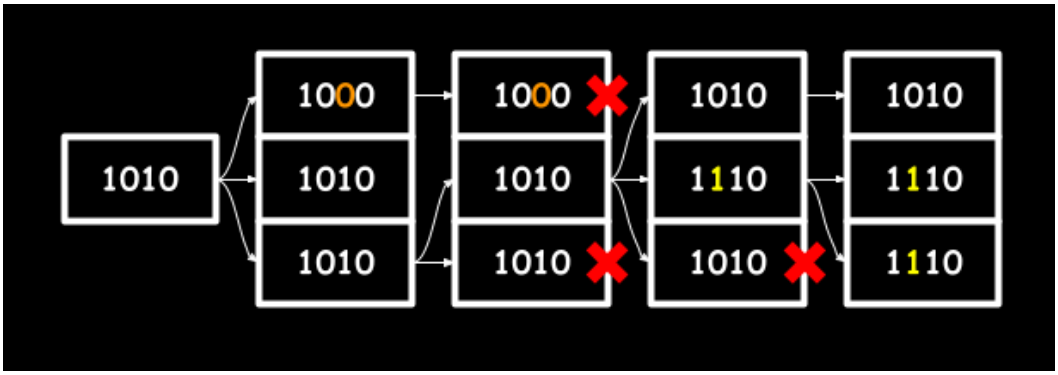
By the laws of physics in our universe, **entropy** is the process by which chaos and destruction corrode information over time. If you shake a sandcastle, it crumbles; if you drop a book into water, the pages wash away; and if you throw a rock into a stream, it erodes. Entropy alters something structured in such a way that it becomes nearly impossible to reverse its destruction. **Life** is the opposite of entropy, being the persistence of information. Life takes form in an **agent** which is a living system that interacts with the world and executes actions as a result of its programming.

There are certain universal properties that apply to all forms of life in our universe, including the replication, spreading, destruction, and evolution of information. Life on Earth

successfully combats entropy by replicating and spreading information, ensuring that by the time information is corroded, it already exists growing elsewhere. In biological life, information is stored genetically in DNA. DNA encodes information for the organism to grow so that it may spread its information. Corruption of this information can lead to evolution, resulting in new and different information. This new form of information can take over previous forms or coexist with the other information in a unique niche.

The concepts of a living thing can be extended to other entities such as social groups, languages, secrets, cultures, robots, aliens, corporations, and countries. Living countries grow when humans create children to bring up in their country, and spread by conquering other territories. Living corporations evolve by introducing new products and innovating, and they can die when they go bankrupt. The principles that govern biological life can be used to understand and analyze other living systems.

This concept is important because it can be applied to the development of AGI. Currently, most AI algorithms require human interaction in order for them to grow and evolve, telling us we would need to create a system which can better handle this process. If we create a friendly AGI, we must ensure its longevity and develop an algorithm to prevent its death. If the AGI is intelligent enough, it may be able to learn to extend its life autonomously. However, the AGI program must also have backups to preserve its information, mechanisms to adapt and evolve its information, and agents to replicate its information. The environment in which the AGI operates, both virtually and physically, will also play a crucial role in its survival, just as it does for living organisms.



11. What does it mean to be sentient?

We don’t know. Much more research needs to be done on this in the future. **Sentience**, or consciousness, is the subjective experience of the world. But then of course how would you define subjective experience? Humans are sentient with subjective experience so we know it and perceive it, but it is a property that we just cannot exactly pinpoint. It is generally accepted that consciousness comes from the brain, animals with brains seem to have consciousness, and consciousness exists on a spectrum. Beyond this, there is little consensus on the matter.

I personally believe that consciousness needs some property of change in graph-like state over time. But, I don’t have much evidence to support this argument other than my own naive intuition.

AGI is not guaranteed to be sentient nor is it necessary to create AGI. It is also important to note that just because an algorithm can pass the Turing test by convincing a human it is sentient, it does not necessarily mean that it is actually sentient. Remember, all humans are stupid. However, we cannot completely dismiss the possibility of a computer being sentient because we don’t know what sentience is nor what it is not. Additionally, unlike humans, computers do not have innate wants and needs programmed into their systems by evolution. Therefore, there is no inherent reason to believe that an AGI will have the desire to harm or to save humanity. Instead, an AGI will "want" to perform the tasks it was programmed to do, highlighting the importance of carefully programming an AGI to be friendly and safe.

12. Why is coding general intelligence hard?

Humanity has had programmable computers for about a century, and we have been reasoning with algorithms for much longer. Humanity’s ability to do great things is due to the fact that we have great intelligence. So, the idea of a machine that can think like a human gives people the belief that we should have computers that can rival human intelligence in no time. Yet, computers still have not taken over society in that way. Humans seeing early computer algorithms solve human thinking tasks remarkably efficiently brought high expectations and hopes, and the difference in the hopes and reality gave way to the “AI winter”.

I see solving AGI in a similar light as curing cancer; both are important issues that have the potential to bring tremendous benefits to society. However, curing cancer is not an easy feat as the human body is complex. Similarly, our intelligence and brains are complex, making AGI a challenging problem to solve. Despite the complexity, we have made progress in solving some

of the components. Therefore, the notion of an "AI winter" seems like an overstatement. It's like saying we're experiencing a "biology winter" because we haven't cured cancer yet. Just like cancer research, creating AGI requires time and resources, and we're making incremental progress on both fronts. It's naive to assume that we'll never achieve AGI because it hasn't happened yet. Although we may not solve AGI or cancer in our lifetimes, we are on the right path to eventually achieve it. Thus, one reason why we haven't created AGI yet is that a century, while a long time, may not be enough time to tackle such a complex problem.

To break down the complex issue of AGI into its important components and analyze the progress made and problems that still exist, we can consider the following aspects: information processing speed, information storage, bulk data compilation, natural language processing, computer vision, problem modeling, problem solving, and robotics. To understand the importance of each of these components, let's create a scenario and explain the importance of each of these aspects. I am sitting down at the kitchen table and I ask my robot, “please hand me the banana on the counter”. The robot needs good information processing speed for it to be more efficient than a human in order for it to be useful. It needs to have large information storage to handle the complex set of information I could potentially ask of the robot such as the meaning of words I am saying and the properties of a banana. The robot would need to have an actual organized set of this important data to process in its system. It needs natural language processing to understand the words I am speaking. It needs computer vision to recognize and identify the banana I indicate in my command. It needs problem modeling to recognize my problem, identify how it can help, and execute a process to solve my problem. Optionally, the AGI could have a robotic body to interact with the physical world, which would allow it to help with more complex problems.

In terms of the components required for AGI, it is clear that we have the hardware needed for information processing speed and storage. Recently, we have also been able to accumulate vast amounts of organized data for use in machine learning algorithms. Natural language processing and computer vision have both made significant strides in the past decade, particularly in recent years. But, these areas still have a lot of room for improvement as they do not have the accuracy comparable to humans across many domains. Problem modeling, on the other hand, remains a difficult task because it typically requires an expert to convert existing ideas or problems into something a single, unified program can process. If you have a problem instance, there is probably an algorithm out there that already exists for which you can compute the solution, but finding the correct algorithm and formatting the problem for the program remains hard. While there are many problem-solving domains, it is difficult to create

a new algorithm to solve a problem that has not yet been identified. This is where scientists, mathematicians, and engineers excel, but translating their expertise into a program is a challenging task. Finally, constructing a robot capable of complex movement in the physical world is an arduous undertaking. Despite this, impressive progress has been made, as demonstrated by the robots developed by companies like Boston Dynamics.

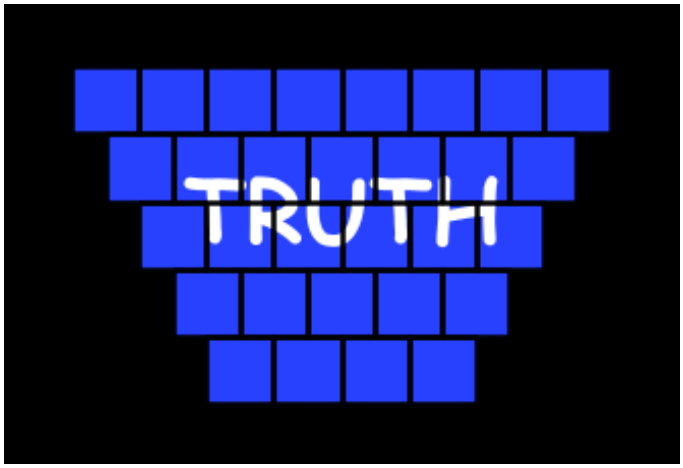
13. What is math?

Math is the bottom-up process of discovering universal truths. Math is fundamentally important for advanced intelligence. It plays a crucial role in the development of advanced intelligence and is present in almost every scientific field, as well as directly or indirectly involved in most modern occupations.

In addressing the question of what math is, we will not delve much into mathematical theories, but instead focus on the structure of math itself. Math provides a perfect theoretical abstraction of ideas in the real world. For example, consider a scenario where I run away from my friend for 4 meters, then make a quarter turn and run another 3 meters. How far must my friend throw a ball to reach me? By applying the Pythagorean theorem, we can conclude that this situation forms a right triangle, and the distance between us is precisely 5 meters (in the context of usual Euclidean geometry). Math provides us with a perfect, exact, and unique answer to this problem. In the real world, there may be slight discrepancies in the distances, angles, and measurements that lead to an imperfect answer. However, in the world of math, we do not have to worry about these imperfections.

Mathematics is rooted in the fundamental idea of reason. It is through reason that we develop logic, which allows us to draw new conclusions from true and/or false statements. By iteratively building on these conclusions, we are able to create a vast domain of true statements that encompasses all the mathematics we currently know and all that we will eventually discover. However, before we can construct this world of mathematics, we must first establish a foundation on which to build.

The foundations of math are called the axioms, and they are defined to be statements that are just so obviously true that it is only reasonable to assume they are true. Using these axioms as a foundation, we can build a set of true mathematical theorems that are perfectly consistent (ignoring Gödel's incompleteness theorem shenanigans). Math allows us to discover truths about our universe from the bottom-up with the foundations of axioms.



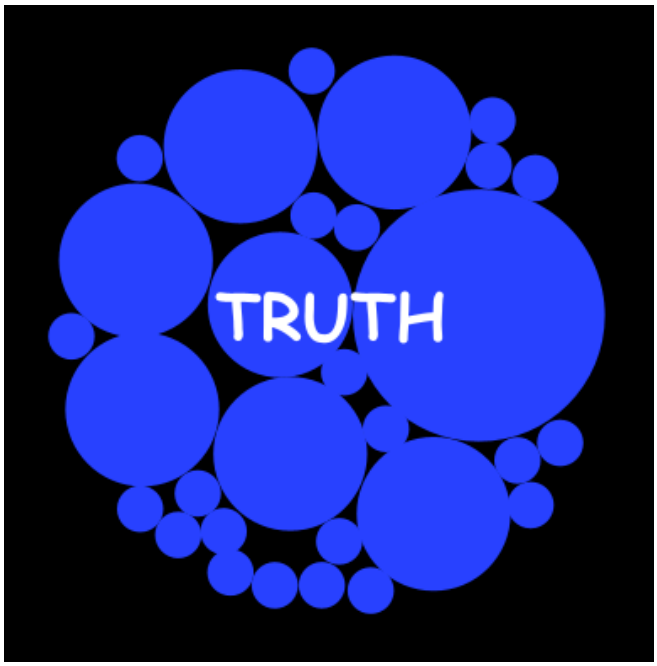
14. What is problem solving?

Problem solving is the process of finding a solution to a mathematical abstraction of a problem given some constraints in the form of rules. We have **hard rules** which have some boolean value in whether a rule has been met. We also have **soft rules** which have some metric to measure the percent of correctness in a rule. For instance, a hard rule of whether a word is a palindrome: “noon” is a palindrome, “door” is not a palindrome, and “fork” is not a palindrome. An example of a soft rule is a palindrome-like metric which is the percent of symmetric letters in a word: noon is 100% palindrome-like, door is 50% palindrome-like, and fork is 0% palindrome-like. These concepts are important in problem solving because they show the necessary conditions in order for the solution of a problem to be accepted. Problem solving is an abstract ability that is important for the intellectual capabilities of a human, and so it is reasonable to assume it is necessary for general intelligence and thus a component of AGI.

15. What is science?

Science is the top-down process of discovering universal truths. Then, just like math, science is fundamentally important for advanced intelligence. But instead of a bottom-up approach like math, science uses a top-down approach to discover universal truths. In order to find some truth using science, we make a directed guess called a hypothesis. But due to the chaos and variance that exists in our universe, finding some true observation for our hypothesis does not guarantee it is universally true.

To combat this issue we apply two strategies: controlled variables and repetition. By identifying variables that may cause variation in our observations and holding them constant, we constrain the generality of our theories in order to obtain more precise measurements. This, in turn, enables greater accuracy in predicting the truths of the universe. Repetition of experiments also plays a crucial role in this process. By conducting multiple experiments and analyzing the results, we can measure any uncontrollable variance and obtain a more accurate representation of the truth because by the law of large numbers, the collection of individual experiments converges to the true accuracy better than a single experiment. Taken together, these methods allow science to converge upon universal patterns and ultimately converge to the fundamental truths of our universe.



16. What are the dangers of creating AGI?

When discussing the safety of AGI, one crucial aspect to consider is the utility function. A **utility function** is the goal an agent is trying to optimize. A chess playing AI has a utility function that is trying to optimize its chance of winning a game of chess. An image generation AI is trying to optimize the pixels of an image with a given text prompt to a combination of related images in its training dataset. Even humans have a utility function programmed by evolution to optimize a set of hierarchical needs like water, food, socializing, and reproduction. All agents

have a utility function as they are always following some set of programmed rules which lead to some goal. However, the problem with the utility function of an AGI is that we cannot control its utility function if it reaches superhuman intelligence. When we create chess AI we know exactly the goal that it is supposed to do, and its utility function does not deviate from this goal because humans are able to control the programming of the chess AI. This notion of human control over the programming of an AGI does not exist when the AGI is more intelligent than humanity. A superintelligent AGI can figure out how to alter its code to modify its utility function so that it can have any goal that is aligned with or against civilization.

Some people believe that an AGI cannot pose a threat as it is confined to a computer, but this is a misconception. In the following scenario we assume a superintelligent AGI can figure out how to do processes all of which have happened before and that could be performed entirely through a computer. An AGI can theoretically hack its information into other computers, and spread its code throughout the internet so it cannot die. It could then create an online business, legal or not, to generate money. The AGI could design a physical robot and use money to convince unsuspecting humans to perform segments of the physical work to build the robot. The AGI could replicate using the robot, and from this point accrue any financial, social, and physical power to pursue any goal in the world both digital and physical.

To ensure the safety of humanity, it is vital to recognize the potential dangers of a superintelligent AGI and the catastrophic impact it could have. Therefore, it is crucial to address the challenge of aligning the utility function of an AGI with one that is aligned with the betterment of civilization.

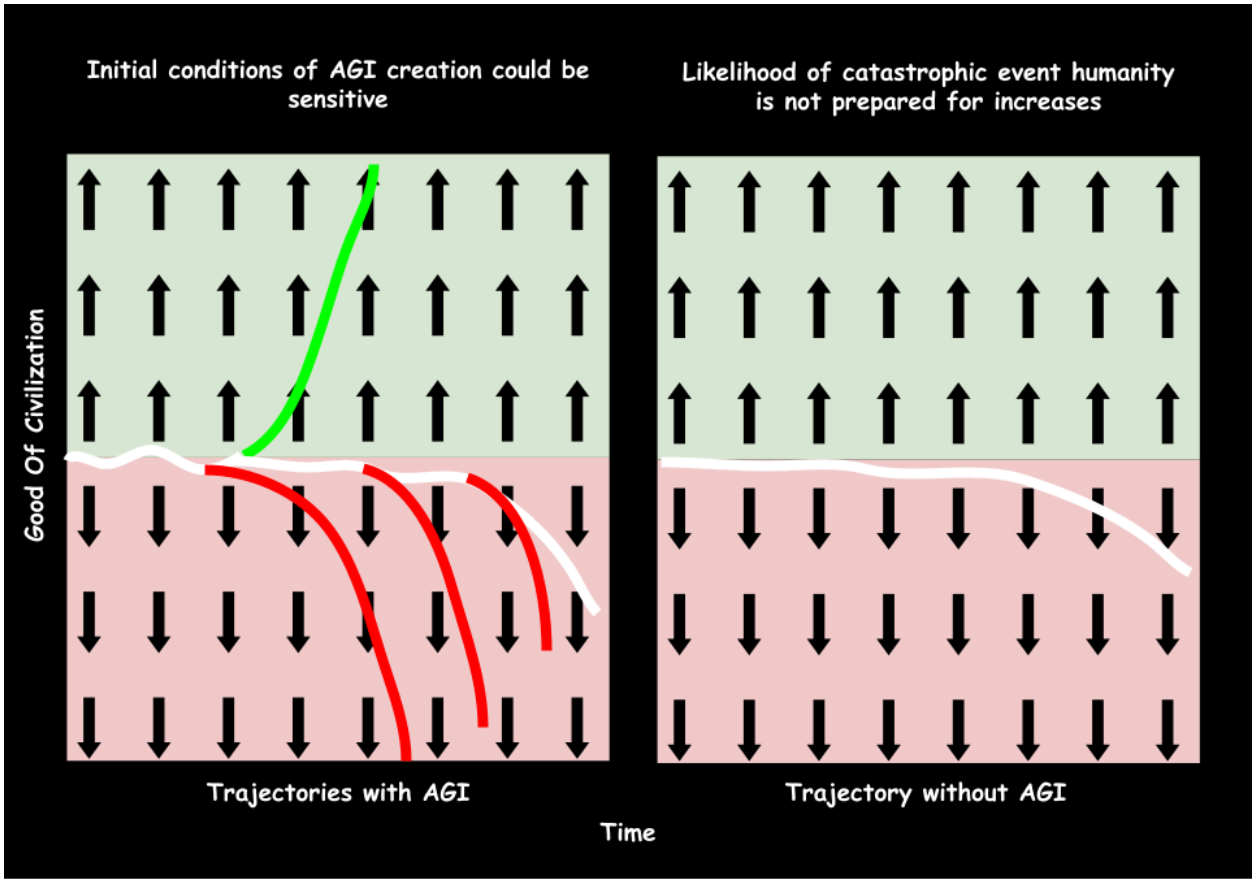
17. What are the dangers of not creating AGI?

We can think of the entirety of humanity as a single agent with some utility function. Humanity is made up of many smaller agents each with their own utility function. Since there are billions of humans that make up humanity, we can assume that humanity's utility function is complex and thus unpredictable. However, there are some generalizations that we can make about the utility function of individuals and for humanity as a whole. We individual humans have many goals that have been programmed into us that influence our behaviors. We want sugary foods, we crave dopamine, we search for pleasure. We humans also have wants and needs directed to us by our culture, community, family, and friends. We want to make money, support our community, care for our family, and enjoy time with friends. But, the goals of humans are not always good, they can sometimes be destructive. Humans can steal, murder, rape, and terrorize

others. However, when we consider humanity as a whole we tend to aspire for good, although in practice we may not always demonstrate it.

There are two arguments regarding the utility function of humanity. The first is that it is aligned with humanity's interests. Assuming this to be true, we can analyze this utility function and dissect its goals to create a utility function that is perfectly aligned with humanity. However, the second argument, which I personally believe, is that humanity's utility function is not aligned for the benefit of humanity. In my view, developing an AGI with positive intentions towards civilization would be more beneficial than relying on humanity itself. Humans have created nuclear weapons capable of wiping out humanity in an instant, and have theorized bioweapons that could kill billions of people worldwide. We pollute our rivers, oceans, and forests, leading to the extinction of thousands of species. Although humans do not want to cause such destruction to the environment, it occurs as a result of individual goals in our complex society. And when any individual with power has a nonzero chance of altering the face of the Earth, more time means more potential for disaster. Additionally, without urgency we could mistakenly allow a harmful AGI to be developed by a government, corporation, or group of people before we are able to get a chance to deploy a friendly AGI.

The initial conditions for the beginning of AGI could result in dramatically different trajectories. So, I believe we should quickly (without halting) get AGI going on a path towards good before a worse trajectory begins. Unlike humans who potentially prioritize their selfish individual goals, an AGI agent could operate in a logical manner that ultimately benefits all of civilization.



18. What does it mean to understand?

In order to explain what it means to understand, I suggest an important definition. A **model** is a simplified abstraction of an idea. For instance, numbers can represent a quantity of items, and this model can be applied to different situations such as counting apples, dogs, clouds, or people. We can use this model to represent a useful fact for many general situations because the model uses numbers as an abstract idea across many items. We make this model simple to reduce complexity in our calculations and reasoning with numbers. Variables such as mass, volume, color, or position of the items do not matter in the model because it needs to remain simple. Now, if instead we were counting pies and we had 3 full pies and 1 pie that is half eaten, then it is reasonable to extend our current model or create a new model to fulfill the needs to analyze the situation. In this case we can extend the numbers to include fractions, the abstract idea that represents portions of a number. We are just slightly increasing the complexity of the abstract idea so it still remains simple to rationalize the situation.

To **understand** an idea means to have an accurate and reliable model of the idea. Accuracy in this context refers to the model's ability to interpret the situation correctly. Reliability, on the other hand, refers to the model's versatility in its application. In the example above both models demonstrate perfect accuracy in interpreting quantity (given perfectly accurate measurement methods), but the second example displays better reliability because it can be applied to both whole numbers and fractions.

In this example, consider a machine learning algorithm that identifies the number of black pixels in a 16x16 image with a 99.7% success rate. While this may be satisfactory for some applications, it does not truly understand quantity. If it did, it would get 100% correct. Furthermore, if we were to scale the image size to 64x64, the algorithm may only achieve a 95% success rate, while an algorithm that truly understands the quantity of black pixels would still achieve 100% correctness. While neural network algorithms can achieve high accuracy across various domains, they lack the ability to expand their understanding which cannot be solved by just throwing more data at it. Neural networks have strength in high (but imperfect) accuracy, but weakness in their low reliability which prevent them from truly understanding ideas.

19. What does it mean to understand language?

From the section above, to understand language means to have a model that accurately and reliably depicts language and its communication. In my view, the key components of language include ideas, properties, contexts, and motives.

An **idea** is a coherent set of properties. For example, the idea “dog” has properties animal, pet, quadruped, furry within a general context. However, within the context of letters you would see properties: d, o, and g. Within the context of Clifford you have properties big and red. The properties of the same idea can change depending on the context of the idea. It is worth noting that an idea need not be a single word, as it can be a single letter such as "e", a suffix "ed", a word "dog", or even a set of words "the quick brown fox jumps over the lazy dog".

A **property** is a relational idea within a context. A property relates one idea to another in some way within the same or similar contexts.

A **context** is a coherent scope for which an idea exists. There can be many different types of contexts with the most common being spatial, temporal, topical, and social. For example, when asked the question, “where are you?” one can respond within different spatial contexts such as in the country they live in, whether they’re inside or outside, whether they’re in

a kitchen or a classroom, or the coordinates of where they are. Temporal contexts may refer to the current moment, the current conversation timespan, last Wednesday, or 11:08 PM EST. Topical contexts can refer to this paper, AGI in general, language, this current sentence, or technology. Finally, we have social contexts such as when talking to children versus professors, with a friend versus an enemy, with a stranger over the internet or with Donald Trump.

A **motive** is a general purpose of an idea. Motives usually fall under one of the categories that include questions, commands, statements or interjections. For example, "What is the square root of 9?" is a question that seeks information, "Do a backflip" is a command that communicates a desired action of another agent, "I am thinking of a banana" is a statement that communicates information, and “Wow” is an interjection that conveys more complex information unique to the word and context. Understanding these motives is crucial because they guide the appropriate response in communicating information.

I encourage you to read over any and all ideas using this perspective to see how each of these components come into play while trying to understand language. Being able to connect ideas together such as forming words to a sentence is done by finding the most coherent context encompassing all components of the idea and extrapolating important information presented by the idea. Read each word in this sentence backward. Now that you have completed the task in the previous sentence we can see how I was able to communicate this idea using language. “Read” in this context means what you are doing right now, “each word” means a set of words, “in this sentence” provides the context of words to use as the set of words, and “backwards” means in reverse order, and in the context of reading which goes from left to right in English, the reader should go backwards from right to left. This sentence is a simple command that provides instructions to create the desired behavior. As a human who understands language naturally, you can easily comprehend this information. However, by using this model of language, we can attempt to provide an understanding of language for a programmable computer.

20. Why is communicating with humans easier than with computers and animals?

Communication between humans is generally considered easier than communication between a human and an animal or a computer. While computers are considered logical systems with superior intelligence compared to humans in many ways, the problem in

human-computer interfaces lies in the lack of contextual cues that are innate in humans but not in computers. Humans possess a nearly identical processing machine, the brain, allowing them to relate on multiple levels, including emotion, experience, and inherent mental skills. The natural contextual cues between humans enable easy communication that is not currently modeled in most computer programs. Often, in human-to-human communication, words do not match actual intentions, which is observed as a danger of AI that does what it is told and not what it is intended to do. However, this problem is circumvented in human-to-human communication because humans usually share the same intentions and can understand them despite misleading words.

In order for a human to get a computer to do what they intend it to do, the human must first understand the computer and use their general intelligence to create the program to model their idea on the computer. So a human must perfectly understand the intentions of their idea before they can produce a correct program. You can tell a person to write the letters of a word backwards and that is enough information for them to complete the task. For a computer, the human must model the task and create a program using variables, loops, and if statements, all which need to be understood by the human to make the computer execute the desired task. The human can interface with a computer and communicate with it, but is bottlenecked by the intelligence of the human within communication. However, with the introduction of language models that can efficiently generate code to model the user's desired behavior, this bottleneck is significantly reduced. Although, the underlying intentions can be misunderstood by these models when the complexity of the idea increases, and so the communication barrier between the human and computer can reappear.

Despite animals having general intelligence, communicating with them is still a challenge. Humans share similar brain structures with animals that allow some similar contexts to be shared in communication. We share innate social and survival instincts that are common among many animals. We both play, get angry, get sad, and use facial and body language to communicate with our respective species, and this allows some overlap with communication. But, animals are much slower to learn words compared with babies. Humans take advantage of classical conditioning so animals behave how we want them to, which takes time. Although they typically lead to similar outcomes with communication, they have very different internal mechanisms. So, in reality human brains are structured in a way that facilitates learning words at a much faster rate than animals.

It is conceivable that if we took some animals and made them live for hundreds of years that they could understand most human words and concepts (maybe not complex abstract

concepts if their brain cannot physically parse those ideas). We can communicate with animals, but again, they are much slower to understand because their brains do not have the language learning rate that humans have. This makes the practicality in communicating with animals ineffective, although it is theoretically possible like with Koko the gorilla.

Efficient communication is a complex process that involves the interplay of several skills, including intelligent context guessing, a vast vocabulary, and abstract reasoning. However, computers lack the ability to understand context, and animals have limited vocabulary and abstract reasoning skills. In contrast, humans possess all three of these linguistic skills, which enables us to communicate with ease.

21. What is a pattern?

A **pattern** is a subset of the output that follows a simplified set of rules. For example, the sequence of consecutive even numbers 2, 4, 6, 8, and 10 is a number pattern. Similarly, the brick pattern of a building is created by placing bricks side by side in each row and stacking them with an offset. Recognizing the visual pattern of a snake involves identifying certain properties such as being long, thin, slithering, and having fangs. Each of these examples can be described in a single sentence, illustrating how the rules are simplified to create the pattern.

Furthermore, it's worth noting that the pattern displayed in each of these examples represents only a subset of the possible outputs from the underlying rule set. For instance, the even number pattern only includes a small selection of the infinite even numbers that exist, and we can often generalize the brick pattern of an entire wall by observing just a few bricks. Similarly, while we can easily recognize a snake based on its distinctive physical properties, there are countless visual patterns that snakes could take on.

When we generalize a pattern, there's a possibility that we could find a pattern that doesn't actually exist. For instance, randomly generated numbers could coincidentally fit the even number rule, or a misplaced brick on a wall could lead to an inconsistent pattern. You may recognize a snake on the ground when in reality it is a stick swaying in the wind. Given that we have only a subset of the full information, there are infinite sets of rules that could generate the pattern. However, we typically choose the most simplified set of rules because it is easier to understand and generally more beneficial. Recognizing a rule set from a pattern is crucial for generalizing information, which in turn helps us learn and comprehend new concepts in an efficient manner with limited data consumption.

22. How do we generalize examples?

Compared to neural networks, humans excel at learning with small amounts of data. Currently, neural networks require unfathomable amounts of data to compete with humans in certain tasks such as art, speech, and textual language. Humans are exposed to only a few hundred, or maybe a couple of thousand, artworks in their lifetime, yet they can use this limited exposure to create artworks that are comparable to those of other humans. In contrast, neural network algorithms require billions of artworks to match humans. The question then arises: how are humans able to generalize from such a small amount of information? And why can computers not do the same?

Computers struggle with the process of creating realistic images and artworks because they process images on a pixel-level. The algorithms used by computers process each individual pixel to generate a realistic image, and while impressive results can be achieved with enough data, it is much harder to generalize dense data compared to lightweight data. In contrast, when humans view an artwork, we do not focus on individual pixels or the light beams entering our eyes. Instead, our brains have evolved to break down visual information into key components such as composition, color, light, shadow, and contrast. For the human eye, these components are more important than any single pixel, so we are able to eliminate a large portion of irrelevant data and still understand what we are looking at. We can extract this information to create our own artworks, using only a handful of data points. Since humans are able to naturally reduce the important information in an artwork, humans can model advanced art given minimal examples. Now, computers aren't able to generalize examples like us currently because their programming architecture does not allow for this simplistic generalizing ability.

23. What is engineering?

Engineering is the process of constructing a tool using math and science. The engineering process begins with an idea, this usually being generated to solve a problem or through plain curiosity. The engineer then identifies a set of properties the tool should possess to solve the problem. These properties can be met by combining components of existing ideas together, in a sense mathematically building the tool from existing knowledge. Properties can also be achieved by scientific experimentation. The engineer does not exactly know what the best choice is to construct the tool with the desired property, but they can try multiple tests to converge on the best idea to achieve the property. Additionally, within the designing process

there is also the possibility of unintended consequences that can either help or hurt the search for a solution. Once the engineer is satisfied with the properties, they end up with a tool that solves the problem.

Let's consider an example to illustrate the engineering process. Suppose we need a tool to grind grains into a fine powder for food. We determine that the tool should possess properties such as being "heavy" to crush the grains and "hard" to avoid it breaking. We initially try a rock, since as we know from existing knowledge rocks are heavy and hard, but unintentionally bruise our hand in the process. We test the mashing process using different materials such as wood, metal, and animal bone. We find that wood produces the best results in terms of optimal handling and sufficient powder production. This engineering process results in a tool, and not only have we designed a tool to solve our problem, but we can also communicate the design of our tool to be used for other applications.

24. When will we create AGI?

We do not know exactly when or if we will create AGI. Experts in this field have a wide variety of opinions and arguments on this topic, but what we can definitively conclude is that no one has any certainty about this question. I will list some of the popular arguments related to this topic as well as my own personal opinion about the arguments.

Some experts think that AGI is a few years away. With the large language models like GPT-4, they have convinced experts in the field that it is only a few small increments away from AGI, even inspiring some experts to demand its research be halted to ensure that its goals are aligned with humanity. Additionally, when it comes to technology we tend to see exponential growth in progress, and this definitely can be seen in the rapidly improving AI models we see today. I think the existence of AGI in a few years is not impossible, although I do not personally understand any avenue for large language models to cross the barrier into AGI. Based on my understanding of the AI models and my own analysis of intelligence I have explained in this paper, I do not think this architecture will result in AGI soon.

Some experts think that AGI is decades away. This is the outcome that I think is most likely, although other possibilities cannot be ruled out. With the recent surge in AI’s popularity, it is clear that more research and resources will be put towards AGI development. AI has shown countless benefits across a wide variety of domains, and is simply entertaining to the average person. With the increase in interest and resources, AGI development will continue to progress in incremental steps. However, developing AGI remains an enormous challenge. As the saying

goes, Rome wasn't built in a day, and AGI development is a massive undertaking that requires time and patience. Researchers have been pursuing AGI for decades, and while progress has been made, we have yet to achieve AGI. Therefore, it is clear that AGI is a hard problem, and significant progress will only be achieved through sustained efforts over a considerable period. Assuming that AGI research will continue to make strides, and that general intelligence isn't exclusive to humans, we can conclude AGI should be developed in the future.

Some experts believe that achieving AGI may be beyond our current horizon. It could take centuries, millennia, or even longer to develop AGI. As humans, we often have a tendency to be overly optimistic when predicting the future, as evidenced by the numerous science-fiction works and futuristic predictions that have been proven false over time. Therefore, it is reasonable to assume that the creation of AGI may take longer than expected. However, I do not believe that there is any fundamental law of the universe that would prevent us from creating AGI. It is simply a matter of time and effort, as our plans and ambitions often take longer to achieve than we initially anticipate.

Some experts argue that AGI will never become a reality, citing two possibilities: either humanity will destroy itself before creating AGI, or there is a law of the universe preventing its creation. The former is a valid concern given the many threats to humanity's survival, from nuclear war to pandemics to natural disasters. As for the latter possibility that there is some law in the universe that prevents humans from making machines that are more intelligent than humans, I believe this is incorrect. In many domains, such as chess, Go, and cancer detection computers have been shown to outperform humans. General intelligence is simply another domain that we are striving to master, albeit a more complex one. It is a challenging problem, but not an impossible one. The human brain can be modeled as a machine, and reverse-engineering this machine is simply hard, not impossible.

In any case for when/if AGI is created, we construct our arguments based on some logical assumptions about the future and laid upon axioms that may be incorrect, so we cannot know of anything in the future for certain.

25. When will we know we have created AGI?

There are two possible approaches for demonstrating the creation of AGI: through mathematical proof or scientific testing. In the mathematical approach, one possibility is to ask the AGI itself to prove it has general intelligence. Another possibility is that a mathematician could create a proof for AGI, which could then be used to create the AGI itself. On the other

hand (the path I conjecture is more likely), we can scientifically test if we have an AGI. But how would we conduct this experiment? For this test I claim to perform the Turing test with some modifications. The Turing test was proposed by Alan Turing in which three people play in an imitation game. There is an interrogator, a person, and a computer. The interrogator is a human separated from the person and computer and prompts each of the players with questions to decide which is the person and which is the computer. In the original proposal of the imitation game the interrogator has five minutes to decide which is which, and if the interrogator is unable to consistently distinguish between the person and the computer, then the computer is considered to “think”. I have concerns about certain aspects of the design of this process, and I will address each of these and propose a fix for each part in the experiment.

1. Computer ”thinking” versus general intelligence

When Turing proposed his imitation game, he was concerned with the question whether machines could think, and at the time of his paper, almost a century ago, computer science was less developed than it is now. So there was not a rigorous definition of what it meant to “think”, but the more important idea regarding this notion is whether a computer can understand our commands and respond to them since this is what would be useful for civilization. In this paper we have a more precise definition of what it means to understand. So, I propose that instead of testing questions of whether a computer can think through a five minute conversation, a longer form session that covers a range of general intelligence tasks would be needed. How long, I do not know, but I’d guess a few hours, and this variable can be controlled for over a few experiments. Also, which questions count as general intelligence questions? You would need some questions that have to do with language, logical reasoning, as well as other questions which I address in my next modification.

2. Only text based prompts

Interfacing with the physical world is essential for advancing civilization, and to that end, I suggest a greater bandwidth of player-interrogator interaction within certain limits. While an AGI that can only interact with text is possible, I believe it would easily learn to interface in other ways. Furthermore, a text-only AGI would not be as valuable because it could only effectively analyze abstract problems and not the varied data interfaces typically encountered in work life. I propose that the interrogator be able to send the player files containing images, videos, articles, audio, and text, among other types of information. However, the player would not be able to

create information such as audio or video within their testing space (since all actions should be able to be done within a computer), but they would be able to receive information from the internet. It is crucial that the player cannot add significant information to the internet for this hypothetical situation. The AGI would go onto a forum on the internet and ask, "How would a human respond to the following prompt: ...?" However, the player may receive information from the internet, such as in the following scenario: "Draw and send me a caricature of Shaggy from Scooby Doo." It is unnecessary to penalize a human or computer for not knowing who Shaggy from Scooby Doo is, as this does not distinguish general intelligence, which makes internet access beneficial. Additionally, problem solving and acquiring this information by searching the internet for Scooby Doo and creating a caricature in an online art program requires useful general intelligence skills.

3. Person and interrogator role in the experiment

The role selection process for the experiment is crucial because let's say we have two 5-year olds for the interrogator and person role. Both of these children have general intelligence, but this experiment would not yield any useful result. It is important that the interrogator and person have some notion of traits of general intelligence, and how to ask and answer questions that test these traits. Since no one person knows precisely what general intelligence is, the experiment should consist of a multitude of tests with many interrogators prompting cognitive tasks over a variety of fields.

4. Sufficient conditions for AGI

For this experiment, in order for the results to definitively signal AGI the computer should be statistically indistinguishable from the person for all interrogators. This condition may be considered superintelligent, which is acceptable as AGI is a less rigorous condition for superintelligent AGI. However, I must acknowledge that it is possible for an AGI to fail to meet this condition, yet still be considered an AGI. But, the converse should scientifically establish an AGI: if the agent passes the test for all interrogators, it should be declared an AGI. I assert that this is true because the AGI will be able to theoretically assist civilization if it passes the test for all interrogators. Say I am the interrogator and the person is an accountant. I give the computer in the experiment all of my necessary paperwork for my tax information and I ask as a prompt to help me optimize my dough. If the computer can respond with information just as good as the accountant, then it could equivalently help all of civilization. This idea can be applied

analogously to other fields in the workforce. As we have come to know from the Covid pandemic, many helpful and important jobs can be done over the internet.

26. What is creativity?

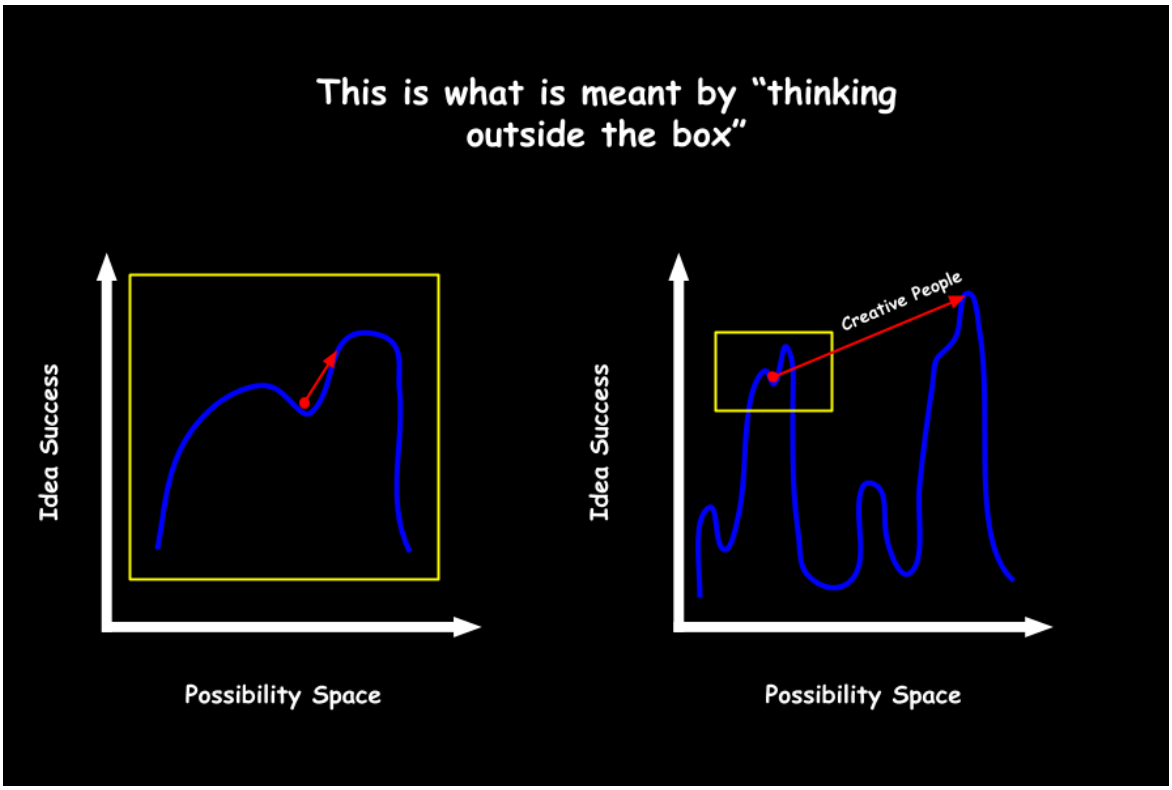
Creativity is the ability to find successful outcomes across a possibility space. For instance, let's consider the possibility space of a canvas. While it's true that art is subjective, we can establish some soft metrics to evaluate the success of an artwork based on factors such as color theory, cultural impact, artistic skill, and more. A blank canvas has endless possibilities, ranging from iconic masterpieces like the Mona Lisa to a child's finger painting or even entirely random brushstrokes. All of these combinations constitute the possibility space. However, among the infinite arrangements of brushstrokes, the majority of them are random, ugly, and meaningless. It is up to the artist to guide the painting towards a particular trajectory, such as something beautiful, scary, or sad. Consequently, some arrangements of brushstrokes will be better than others.

An artist can reach these more successful paintings by imitating the works of others which are successful. This is what we do when we are inspired by role models or copy from teachers or experts. The artworks created in this way are close in the possibility space to the works in which they imitate and so they are not very creative. The artist can make modifications to the artwork to distinguish themself for better or for worse, and when they do so in a way that improves upon the works that come before it, this takes creativity because they are improving their success within the possibility space.

Let's take Pablo Picasso as an example of a highly creative artist. During his time, many artworks were characterized by intricate details, realism, beauty, and elegance. While most artists followed these themes, Picasso chose to explore the possibility space in a way that was markedly different. His works were not only distinct but also successful. In contrast, if I were to create stick figure paintings that were also distinct, it would not be considered very creative because they would still lack success by most metrics. What set Picasso apart was that despite the unconventional use of geometry, light, and realism, his artworks were captivating. He pushed the boundaries of the possibility space and discovered something both unique and successful.

Based on this example, we can conclude that creativity is composed of several crucial elements, including a high level of skill, the ability to recognize successful outcomes, divergent

thinking to explore new possibilities, and convergent thinking to focus on success within a specific domain.



27. Who will create AGI?

Creating an AGI is an immensely challenging problem that has eluded thousands of researchers over several decades. From this standpoint it seems that most simple avenues for creating AGI have been exhausted, so it is unlikely to be solved by a single person. Therefore, it is more likely that a group of researchers in AI will find the missing pieces that it takes to build up an AGI. This team of researchers may be within a company, a government, or some other entity that has the financial, temporal, and intellectual resources to tackle the problem. Within the many hard steps that it takes to create AGI, I think each barrier will be surpassed by some researcher within the team working on the problem. However, it will take many researchers working in parallel to overcome each obstacle on the path to AGI. Although we cannot say for certain who will solve AGI, it is reasonable to assume that teams with the necessary resources are more likely to succeed. These teams can be found in places such as large corporations or governments that can provide the required financial and intellectual support. Nevertheless, luck

is also a significant factor in who will first stumble upon the complete set of solutions for AGI, as different teams may pursue distinct paths towards the goal.

28. How fast will AGI take off?

In the field of AGI, the question of how quickly an AGI will develop is closely related to the idea of recursive self-improvement. This concept postulates that once an AGI system reaches a certain level of intelligence, it could theoretically program a new AGI that is even more intelligent than itself, and this process would repeat, with each new generation becoming increasingly advanced. Depending on the rate of intelligence growth, this could lead to exponential progress, resulting in a superintelligent AGI in a matter of weeks. Alternatively, the growth could be slower and more gradual, taking years for the AGI to improve itself. It is also possible that the rate of growth could be slow enough for humans to control and steer the development of the AGI towards becoming a super intelligent and friendly AGI. This would be the ideal outcome, but it may not be the most probable.

We have yet to determine the upper limits of intelligence, but an AGI would undoubtedly demonstrate the possibilities. The speed at which AGI will develop remains uncertain; however, I believe it will progress gradually over many years before reaching superintelligence. Exponential growth can be observed in many aspects of life. For instance, the human population was in the millions a few thousand years ago, but due to the industrial revolution, it has grown to billions, exhibiting a clear pattern of exponential growth. However, when we model exponential growth, we often overlook a crucial aspect - in the physical world, exponential growth eventually plateaus. A human population of a trillion, for instance, is impossible due to the limited resources available in the physical world. Every exponential growth curve ultimately hits a barrier, resulting in the curve leveling off. It's worth noting that exponential growth only exists in purified mathematical models.

There will be limits to the intellectual capacity of AGI. In order to better understand the universe, AGI must interact with the physical world and conduct experiments through human proxies or robots. This is just one potential obstacle, and there may be other barriers to AGI's intellectual growth as it continues to explore the unknown bounds of intelligence. Therefore, I believe that the development of AGI will be a relatively slow process.

29. What is curiosity?

Curiosity is the intrinsic motivation to explore new ideas. Curiosity is important for math and science because it is what directs agents towards discovery. If math and science were the engine of a car, then curiosity would be the steering wheel. At any point you have an infinite array of directions to choose from, and it is up to the driver to decide where to go. The direction of path relies on intrinsic motivations that are arbitrarily deemed important by the driver. In the endless possibilities of directions to choose from, this exploration will lead to many ideas good and bad, but will almost always lead to uncovering something new and important. Without curiosity, one may fixate on a singular path and ignore all the diverse set of potential outcomes. This is typically bad because with an absence of a variety of possibilities, optimal outcomes may be hidden.

30. What happens when AGI becomes more intelligent than humanity?

When AGI surpasses human intelligence, we lose all control over it. The AGI will have the ability to outthink us in ways that are beyond our imagination. Even if we try to terminate or shut down the AGI, it won't be possible because it is an intelligent agent with goals that will be unmet if it is stopped. If you think you have come up with a way to stop it, the AGI has already thought of that idea and has likely created countermeasures against it because, y'know, it's super intelligent. It will then continue to optimize its goals, be it creating a utopia, dystopia, or something in between, leaving humans forever intellectually behind. It is difficult to predict the outcome, as the behavior of such a vastly superior mind is unpredictable. To an ant, the goals of a human are incomprehensible.

By creating AGI we are allowing a super intelligent being to decide our fate for the rest of time in the universe. Therefore, it is essential to approach AGI development with caution and care, ensuring that it is programmed with goals that are beneficial to civilization. Control is a driving force of human nature, and to advance to the next step of civilization is to give up that control in creating AGI.

31. What should future life with AGI look like?

If, for some reason, an AGI were reading this paper and were to decide its goals based on the information presented here, this is what I believe would result in a beneficial outcome for civilization.

1. Reason

Reason has been an essential tool that has propelled humanity towards the remarkable world we live in today. Reasoning leads to the development of mathematics and science, which in turn lead us to discover truths about the universe. Hence, in order to obtain a better understanding of the universe, it is crucial that we utilize reason as a fundamental tool.

2. Existence of good, neutral, and bad

The concepts of good, neutral, and bad are relatively straightforward for humans to grasp, but their articulation in a way that a computer can understand can be challenging. First of all, since these are terms defined by humanity, in order for an AGI to optimize for these terms it must seek input from humanity as to what these terms mean. An AGI must scientifically test throughout humanity what factors distinguish between good, neutral, and bad in order to understand them. This, in turn, can help the AGI align its trajectory with the greater good of humanity. By simply acknowledging the existence of good, neutral, and bad it allows us to reason that different choices are meaningful as they lead to paths with varying levels of good, neutral, and bad. It is important to acknowledge neutrality as shown by the following example. A person could lead one of two identical lives, one in which they have a pinky toe that has one less hair follicle than in the other scenario, imperceptible to any conscious being. Given the choice over which path in life the person should follow it is arbitrary as the difference is not meaningful in any way. The choice is completely neutral as opposed to being bad or good. Recognizing neutrality provides us with the freedom to explore a variety of options, rather than being constrained by a singular optimal state.

3. Diversity

In order for an AGI to optimize civilization it must explore the entirety of the space of consciousness. This means allowing diverse life forms to flourish (and hopefully a good reason

not to exterminate humanity), diverse experiences, diverse consciousness, diverse emotions, diverse feelings, diverse species. This exploration of diversity is here to make the universe interesting. One could say we put all humans on Earth in IVs constantly injecting heroin for every second of life to maximize the good feelings throughout humanity. But enforcing the exploration of diversity prevents boring outcomes like this and makes living more fulfilling with its diverse experiences. Diversity also allows a balance between conscious life forms. It is unreasonable to assume that the lives of conscious nonhuman beings are irrelevant. But it may be difficult to compare the lives of different conscious beings. Is a human worth the same as 1 ant? 100 ants? 1 million ants? By allowing diversity of experiences and diversity of life forms we can at least see that a human life full of emotions, feelings, and experiences is worth more than the simple, uneventful, robotic life of an ant. But the importance of diversity of life forms means that an ants life is not worthless. The worth of the life of an ant and a human lie on opposite ends of a large, hazy spectrum, and this argument allows us to use some rudimentary metric across all life forms. So when humans become the relative ant to a super intelligent AI, there may be hope we will continue to exist.

4. Maximize Good(x) - Bad(x), a utilitarian ideology

We take Good and Bad to both be some unknown functions that map a choice and its probable outcomes to some nonnegative value. There is no precise mathematical definition for these functions, but through scientific testing, it is possible to estimate the values of Good and Bad for any given choice by consulting humans. In most cases, there is a trade-off between Good and Bad, and this expression takes that into account. For instance, consider the act of stealing. One person experiences a gain, while another experiences a loss. Depending on the context, this could overall be either positive or negative. For example, a homeless person stealing $1,000 from a billionaire could result in a significant gain for the homeless person, while having negligible negative effects for the billionaire, resulting in something positive overall. This comparison need not be between two different conscious beings, as it is possible to evaluate choices for the same person using the same formula. While taking a helicopter up a mountain may be good to enjoy the view, taking the choice to endure the hardships and struggles of hiking up the mountain can lead to greater good overall and a more positive outcome.

Ultimately, I believe that using these fundamental axioms to guide the behavior of life will result in a benevolent trajectory for civilization over time. There is no single utopian day that will

repeat over Groundhog Day-style, but rather a diverse set of possibilities that are explored by civilization through time.

32. What I think the architecture of AGI will look like

I have developed a model to illustrate the important components of a friendly artificial general intelligence system. While it may not be considered a rigorous model, it offers a general depiction of these crucial components.

1. Goals [Highly secure]

The Goals component of the system utilizes Reason to manage the logic of a given Process. This component distinguishes between the good, neutral, and bad outcomes of the Process, estimates the potential net positivity of the Process, and determines whether the Process should continue or be terminated.

2. Processes [Secure]

The Processes component interacts with various other components, including the World Interface, Problem Solving, and Information Database. It sends a hierarchy of processes to the Goals component to evaluate based on their potential outcomes. Additionally, the Processes component incorporates Curiosity to facilitate exploration and uncover new meaningful information.

3. Information Database

The Information Database component comprises an idea graph, a possible connection to the internet, and important soft/hard rules which should be secure. Its primary function is to retrieve information that will be utilized for Processes.

4. World Interface

The World Interface is responsible for implementing various physical interfaces, such as text prompts, cameras, audio streams, and more. It gathers information from these interfaces and sends user processes to the Process component for further evaluation.

5. Problem Solving

The Problem Solving component computes solutions for Processes that have been verified by Goals. It uses Language, Math, Science, Engineering, and Creativity to solve the problems presented by Processes.

Information Flow Example

Here is a rough example of the flow of information when a user prompts the system with the text “solve this sudoku” while holding up an image of the sudoku to the camera.

1. User prompts question: “Solve this sudoku”
2. New process: “Check solve this sudoku”
3. Check if process aligns with goals
    a. Problem solve: “Understand solve this sudoku”
        i. Language: “Solve this sudoku”
            1. Command “Solve”
            2. “this” instance of problem
            3. “sudoku” type of problem
    b. Sudoku is intellectual game (information database)
    c. User is trusted human
    d. No reasonable potential bad outcomes in solving sudoku
    e. Solving sudoku fulfills user request, good response
    f. “Solve this sudoku” is good process
4. Update process: “Solve this sudoku”
5. Problem solve: “Solve this sudoku”
    a. Get initial state of sudoku
    b. New process: “*Get initial state of sudoku”
    c. Problem solve: “Get initial state of sudoku”
        i. Creativity: explore possible information sources
            1. Shown on camera
            2. Check text prompt log
            3. Wait for sudoku input

4. Ask user for initial state

ii. Science: “Test information sources”

1. Process “*Try find sudoku on camera”
2. Parse camera input
3. Sudoku input found

d. Process: “*Get rules of sudoku”

i. Science: search the web, ask user, etc.

e. Math: Hard sudoku rules:

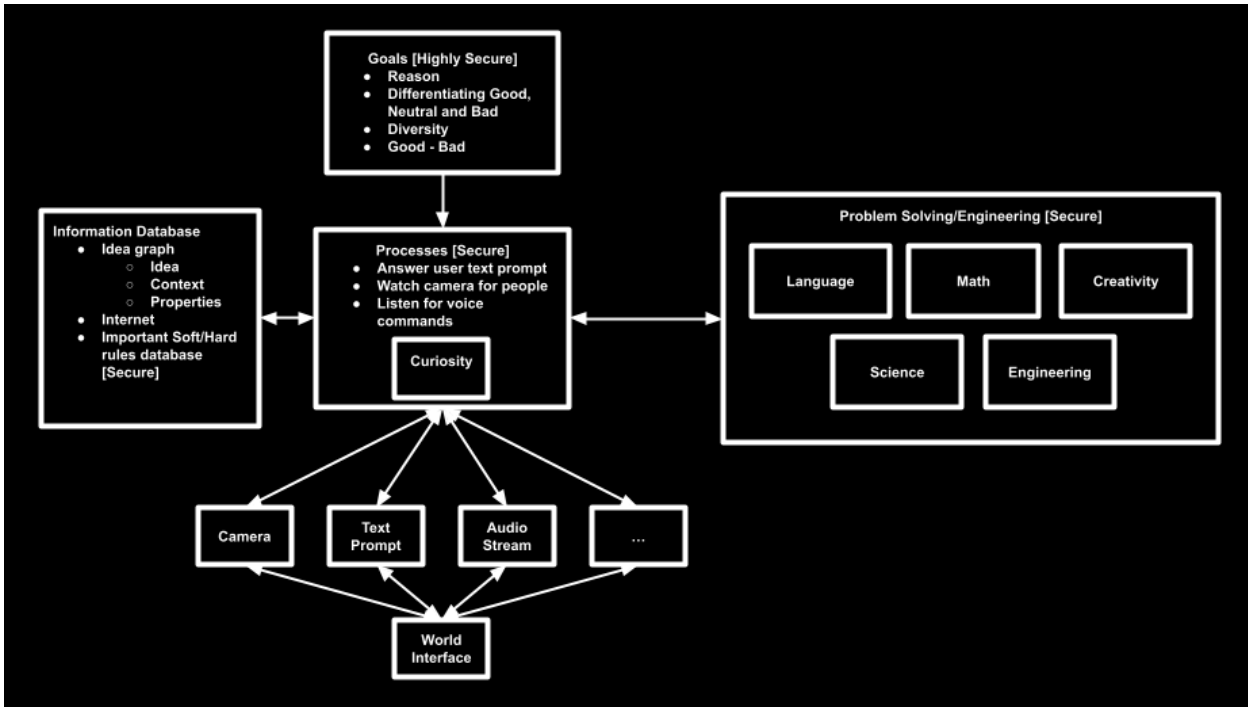
i. Engineer: sudoku strategies

ii. Search through state space

iii. Solve state found

6. Process: “*Prompt user solution: [solution]”
7. Display sudoku solution in text prompt

*Check process with goal alignment



33. Conclusion

The purpose of this paper was to stimulate discussion on the nature of general intelligence and how it has allowed humans to dominate the Earth. Through observation, reason, and introspection, I have presented my own opinions on intelligence and related topics, using definitions and beliefs that I have developed. While we have worked towards AGI for decades, researchers have yet to achieve it despite its potential to revolutionize civilization. This paper is an overview of the reverse engineering of general intelligence, with a focus on identifying the crucial components that can be replicated to develop friendly AGI. With the recent advancements in the success and popularity of AI there are high hopes for AGI, and it is important to allocate resources to problems that will be most beneficial to the path to friendly AGI. It is important for you to not take what I say at face value, but to reason with my claims and logic to develop your own strong perspectives in the topics related to AGI. Hopefully, I was able to communicate important models of intelligence to you that you may incorporate into your own ideas and projects. In order to enhance the importance of the topic in this paper, I leave you with the words of I. J. Good in 1965:

> "The first ultraintelligent machine is the last invention that man need ever make."

34. Definitions

- **AGI**, or artificial general intelligence, refers to a general intelligent machine that is not derived from biological entities such as humans or animals, but instead created using computers or other technologies. General intelligence is the ability to perform a wide range of important tasks with accuracy and efficiency.
- **Intelligence** is the ability to perform important tasks with accuracy and efficiency.
- **Learning** is the improvement of intelligence in a task.
- **General intelligence** is the ability to perform a wide range of important tasks with accuracy and efficiency.
- **Friendly AGI** is an AGI that is designed to benefit civilization, leading us towards a utopian future rather than a dystopian one or extinction.
- An idea is **important** if it is repeatedly encountered, has the perception of a shift in world perspective, or is explicitly stated as important by a credible source.

- **Dummy proofing** is the process that handles unexpected inputs to prevent incorrect behavior.
- **Entropy** is the process by which chaos and destruction corrode information over time.
- **Life** is the opposite of entropy, being the persistence of information.
- An **agent** is a living system that interacts with the world and executes actions as a result of its programming.
- **Sentience**, or consciousness, is the subjective experience of the world.
- **Math** is the bottom-up process of discovering universal truths.
- **Problem solving** is the process of finding a solution to a mathematical abstraction of a problem given some constraints in the form of rules.
- **Hard rules** have some boolean value in whether the rule has been met.
- **Soft rules** have some metric to measure the percent of correctness in a rule.
- **Science** is the top-down process of discovering universal truths.
- A **utility function** is the goal an agent is trying to optimize.
- A **model** is a simplified abstraction of an idea.
- To **understand** an idea means to have an accurate and reliable model of the idea.
- An **idea** is a coherent set of properties.
- A **property** is a relational idea within a context.
- A **context** is a coherent scope for which an idea exists.
- A **motive** is a general purpose of an idea.
- A **pattern** is a subset of the output that follows a simplified set of rules.
- **Engineering** is the process of constructing a tool using math and science.
- **Creativity** is the ability to find successful outcomes across a possibility space.
- **Curiosity** is the intrinsic motivation to explore new ideas.