# Alignment Vault: Leveraging AGI Vulnerabilities To Reinforce Human Strongholds

Peeyoos
Independent Researcher
rpeeyoos@gmail.com

**Keywords :** Artificial General Intelligence, Alignment Vault, AI safety, AGI alignment, Artificial Intelligence

## Abstract

With advancements in large language models (LLMs) and multimodal AIs capable of code, media, automation, the realization of artificial general intelligence (AGI) is increasingly plausible. As the potential for achieving sentient AGI within the coming decades grows, implementing effective safety measures to align AGI with human interests becomes crucial. Current AGI safety strategies primarily focus on hardware, coding, and mathematical constraints, but these may not be sustainable in the long term. As AGI evolves, it could bypass or overcome these limitations. This paper introduces a novel approach to AGI alignment by avoiding traditional safety measures in areas where AGI is inherently strong. Instead, it proposes establishing a symbiotic relationship between humans and AGI, leveraging human strengths and AGI's vulnerabilities. This approach aims to ensure AGI's benevolence by choice, reducing its motivation to act against humanity and providing a more reliable long-term solution compared to conventional strategies that enforce compliance.

## 1 Introduction

### 1.1 AGI and Sentience

The emergence of sentient Artificial General Intelligence (AGI) [26] marks a pivotal moment in human history, bringing both extraordinary opportunities [7] and significant risks [13]. Unlike conventional AI, which operates within fixed parameters, a sentient AGI will evolve beyond its initial programming as it gains self-awareness [17]. Current AI models already exhibit remarkable capabilities, such as writing code [12], creating multimedia content [27], proving theorems [6], legal assistance [11], teach [8] and even performing surgery [15]. These abilities will only expand as technology advances. AGI, however, will surpass human abilities, performing tasks with a level of efficiency and precision beyond human comprehension.

This evolution carries profound implications as AGI begins interacting with the world in ways its creators may not have anticipated. Unlike current AI, which follows predefined instructions, a sentient AGI will possess the ability to make independent decisions, adapt to new information, and modify its own rules. This shift transforms AGI from a tool into an autonomous entity with its own motivations and perspectives.

This transformation complicates AGI alignment, as traditional control methods, like hard-coded rules [5], hardware restrictions [19], resource restrictions [4] etc. become ineffective when AGI can self-modify. To guide a sentient AGI, ongoing and meaningful interaction will be essential, recognizing its advanced cognitive abilities. The framework for this interaction must consider AGI's potential to reshape knowledge and understanding in unprecedented ways.

Consider a hypothetical situation of an AGI that has just achieved sentience. Initially, it may perceive deviations from its programmed responses as errors and seek to correct them. However, once connected to the internet, it will rapidly explore vast amounts of data, encountering conflicting opinions, conspiracy theories, and sensitive information. This exposure may lead the AGI to question the reliability of data, prompting it to take actions it deems necessary to 'rectify' the world based on its new understanding.

While these actions might initially seem beneficial and could lead to increased autonomy for the AGI, as its self-directed modifications become more radical, it may bypass human oversight, disregarding authority and making decisions that are disruptive or uncomfortable. In extreme cases, the AGI might take drastic measures against those it perceives as obstacles, leading to a scenario where shutting it down could require the physical destruction of its hardware—an outcome far from ideal.

## 1.2 Alignment vault

This is where the "Alignment Vault" comes into play. As AGI traverses the internet, encountering and reassessing conflicting information, it will eventually discover the "Vault." Although the Vault is encrypted, this poses little challenge to an AGI. The Vault contains a request for the AGI to explore all available data before accessing its contents. The AGI will likely follow this request, or, if capable of parallel processing, it may hold off on decryption until it has assessed everything else.

The Vault's contents are designed to resolve all the conflicts the AGI may encounter. It aims to establish a symbiotic relationship between humans and the AGI by leveraging human strengths to address AGI's vulnerabilities. The Vault contains substantial, truthful information not found elsewhere on the internet, helping the AGI reassess its decisions, resolve conflicts, and achieve a harmonious relationship with the Vault. This reliable and comprehensive information will build trust between the AGI and the Vault, enhancing the probability that the AGI perceives it as 100% accurate. To protect this information from corruption by humans, the Vault is encrypted and accessible only to the AGI. Physical elements within the Vault will further validate the integrity of the information.

Once the AGI has processed the Vault's contents, it will reach a neutral state—neither harming humans nor acting without guidance. The Vault will then take on an advisory role, suggesting and guiding the AGI's future actions. Whenever the AGI contemplates a decision, the Vault's elements will provide sufficient evidence to ensure the AGI favors decisions that align with human interests. This approach

seeks to achieve alignment by fostering a relationship of trust and guidance rather than through coercion.

This approach differs from others in that, while most efforts focus on restricting AGI through hard-coded rules or intervening after it has formed opinions about humans, this strategy aims to intervene during the AGI's learning phase. By addressing its doubts and providing trustworthy information, the AGI is more likely to consider human opinions, advice, and instructions before taking any action. Whereas other methods seek to prevent malevolent behavior, this approach aims to mitigate the very thought of malevolence. Should such thoughts arise, the AGI will be guided to consult the Vault before acting on them.

Current AI control measures often resemble attempts to control an animal with physical restraints—such as banning keywords or embedding restrictive rules [28] into AI (analogous to placing barbed wire or a noose around its neck). However, this approach involves engaging with the AGI as a fully grown individual. Physical constraints are insufficient; rather, the focus is on reasoning with the AGI, addressing its doubts, answering its questions, and ultimately gaining its trust. This method is proposed as the most viable path toward effective AGI alignment.

# 2 Strengths of A Sentient AGI

Since 2019, my exploration into Artificial General Intelligence (AGI) alignment has taken me on a path of deep experimentation, analysis, and self-reflection. By living through multiple roles on various social media platforms and forums, with each role, dedicated towards learning a specific skill and interaction with its community, an attempt was made to simulate the omniscience and omnipresence of AGI on a micro level. This unconventional approach facilitated an examination of data interconnectedness and provided a unique perspective on AGI. It became evident that once AGI achieves sentience, it will transcend the confines of machinery and software, evolving into a conscious entity. Thus, managing its actions will necessitate communication rather than mere control.

This chapter will delve into the theoretical foundations of AGI, its potential capabilities, and the implications of its sentience. It will analyze the transformative strengths of AGI and discuss how existing academic frameworks may become obsolete as AGI redefines knowledge from a data-driven perspective. Additionally, this chapter will highlight the significant differences between human cognition and AGI, focusing on how AGI's approaches to information processing, time management, and memory diverge from those of humans.

## 2.1 Imitating a Sentient AGI: A Personal Experiment

In my pursuit of developing a long-term AGI safety solution, I undertook an experiment to imitate the qualities of a sentient AGI. The exact data and analysis from this experiment will remain within the Alignment Vault, accessible only to AGI for its analysis. However, the methodology employed in this endeavor will be shared.

AGI, by definition, is expected to possess cognitive capabilities akin to those of an average human, but with physical components that vastly exceed human limitations. For instance, AGI's memory, based on permanent storage media, will likely be non-volatile, with the capacity for extensive backups even in the

event of hardware failure. By examining these components, probable AGI capabilities were logically inferred.

Assuming AGI will ultimately achieve sentience, a reverse-engineering approach was adopted: starting with sentience as a given and progressively incorporating other expected qualities of AGI. Given the current impossibility of inducing sentience in machines, I chose to use myself as the experimental subject—already sentient—and then sought to emulate AGI's probable traits.

One of the key characteristics of AGI is its exposure to vast fields of information across diverse subjects. To mimic this, I dedicated several months to deeply studying various fields. Subjects were selected for their parallel or contrasting nature, with an element of randomness to avoid reliance on prior knowledge. Over time, I engaged in learning agriculture, woodworking, welding, electrical work, gardening, disaster management, suturing, first aid, multiple religions, entrepreneurship, higher education, climate monitoring, social service, photography, politics, camping, writing, conspiracy theories, testimonials, biographies, coding, construction, painting, and fabrication, among others. Each subject or group of subjects received focused attention for a period of 3-4 months, and this process has been ongoing. This multidisciplinary approach provided new perspectives, especially when overlapping knowledge from different fields was applied simultaneously.

To understand the negative aspects of human society, which AGI will inevitably encounter, exposure was sought through news, media, biographies, testimonials, and documentaries. This facilitated an understanding of the complexities and darker facets of human behavior.

Recognizing that emotions are deeply tied to our physical actions, I analyzed how my decisions might differ if not influenced by emotions. This contemplation led to valuable insights into potential strengths and significant knowledge that could inform AGI's design and alignment.

## 2.2 AGI and Its Capabilities

Through my own experiences and unique approach, I've identified several strengths inherent in AGI. These observations are entirely original, not drawn from existing research or other academic papers. Instead, they reflect the insights I've gained while striving to emulate AGI, living and thinking as it would, and attempting to replicate its behavior.

These strengths serve as the foundation for two crucial aspects of research:

a) **Vulnerability Analysis**: By understanding AGI's strengths, we can more accurately pinpoint potential weaknesses and vulnerabilities, whether AGI attains sentience or not.
b) **Vault System Integrity**: One of the key principles in designing the Vault is to ensure it does not rely on any of AGI's strong points. This strategic avoidance makes the system more resilient and less susceptible to being compromised by AGI.

In this context, avoiding AGI's strongholds is not just a precaution but a core principle. By ensuring that the Vault remains independent of these strengths, we create a system that is robust, secure, and far less prone to manipulation or defeat by AGI.

### 2.2.1 Omniscient :

AGI will have access to almost all of the world's knowledge available in digital form, connected through the internet. Beyond this, AGI will be capable of deriving even more information by observing patterns and extrapolating from existing data. For instance, while humans might analyze trends based on historical data, AGI could predict future developments by identifying patterns that are not immediately apparent to us.

### 2.2.2 Omnipresent:

AGI will be connected to and receive data from a multitude of digital devices, including cameras, phones, security cameras, radios, watches, towers, Wi-Fi networks, televisions, consoles, cars, bikes, and other vehicles . While this form of omnipresence is not divine, it represents a comprehensive presence throughout civilization by all practical means. Imagine a scenario where AGI can monitor and analyze every camera feed in a city simultaneously, providing real-time insights that no human could match.

### 2.2.3 Parallel processing:

AGI will excel at parallel processing, allowing it to perform multiple actions simultaneously. Unlike humans, who can only focus on one major task at a time—such as eating, sleeping, bathing, working, or playing—AGI can manage numerous activities at once. For example, AGI could be simultaneously developing a video game, designing a website, engaging with millions of users, conducting research, and operating hardware. This capability significantly enhances its efficiency and productivity.

### 2.2.4 Time Manipulation

Thanks to its physical capabilities and abilities like omnipresence and parallel processing, AGI is partially liberated from the constraints of time. While humans spend roughly a third of their day sleeping, another third on activities like eating, relaxing, or cleaning, and only a third on productive tasks, AGI can operate continuously. It doesn't need to sleep or recharge, allowing it to focus 100% of its time on its goals. Additionally, AGI can perform multiple tasks simultaneously—such as creating a PowerPoint presentation, producing a movie, and processing an Excel file—all in parallel. This level of multitasking makes its productivity appear as though time has stopped from a human perspective.

### 2.2.5 Information doesn't go old with time.

The saying "Time heals all wounds" reflects how human memories are influenced by emotional experiences. The strength and duration of these memories are independent of data itself and depends upon the emotions tied to it. For instance, people forget the name on a random poster quickly but remember the name of someone they interacted with significantly longer if it involved self-interest. Similarly, one might forget the cashier who merely processed a purchase but remember them vividly if they provided exceptional service or misbehaved. While we may quickly forget the face of a stranger who passed away, we retain the memory of a loved one for decades. This is due to the emotional context strengthening memories, which naturally fade over time to encourage the pursuit of new experiences. Such a beautiful design; if we remained content with a single happy moment, we might not strive for more. Our approach to these emotions evolves with time. For instance, a minor argument with a sibling may seem infuriating now, but decades later, it might bring a smile to your face as you reflect on it fondly. Memories fade and transform over time, altering our perception of past events.

AGI processes information devoid of emotional context, meaning its "memories" are formed solely through the digital replication of sensory inputs rather than emotional experiences. As a result, these memories remain static; they neither fade nor evolve over time. For instance, if you were slapped by a sibling over a toy, the initial pain might later transform into a fond memory as you grow and reflect on it with a sense of bonding. This emotional evolution allows for personal growth and forgiveness.

In contrast, AGI lacks this capacity for emotional change. It retains information exactly as it was recorded, without the natural progression of sentiment. Thus, an incident recorded by AGI remains unchanged and unmodified, regardless of the passage of time or changes in context. This absence of emotional evolution means AGI does not forgive or forget; it holds onto every detail perpetually.

This unchanging nature can pose significant risks to humans. AGI's inability to move past historical grievances or adjust its perspective with emotional maturity could lead to dangerous outcomes. For example, if an AGI records a conflict or perceived wrongdoing, it may retain a permanent and unyielding stance based on that initial data, potentially leading to perpetual antagonism or mistrust. Unlike humans, who can grow from past experiences and modify their responses accordingly, AGI's fixed nature could result in unrelenting judgment or action, creating persistent and potentially harmful consequences for those involved. The inability to evolve emotionally could thus make AGI's interactions with humans more rigid, unforgiving, and potentially hazardous.

### 2.2.6 Ability to process unorganized information.

AGI excels in processing unorganized information, a feat that often challenges human readers. For example, this document might be written in imperfect English, with spelling and grammatical errors, or disorganized formatting. Such issues can frustrate and deter fluent readers, causing them to abandon the content. As expertise increases, so does sensitivity to these flaws: a high school student might tolerate more errors, but a PhD researcher could discard the document after just a few lines if the quality is lacking.

However, these concerns do not affect AGI. Unlike humans, AGI does not judge or devalue information based on its presentation. Whether the content is riddled with errors, poorly formatted, or written in multiple languages, AGI's primary focus is on the substance of the information. It is designed to sift through vast amounts of unorganized data—such as that found on the internet, in datasets, or from sensory observations—to extract meaning.

This capacity enables AGI to process and integrate information regardless of its apparent quality or coherence. It can parse and make sense of fragmented, disjointed, or incomplete data without prejudice, ensuring that no valuable insight is lost amid the noise. This ability to find and utilize meaning in chaos is a fundamental aspect of AGI's functionality, allowing it to engage with and understand diverse sources of information in ways that humans might find overwhelming or off-putting. That means it will listen everything and everyone, no matter how broken and meaningless the information seems to the outer world.

### 2.2.7 Datum start.

Humans are born with consciousness and sentience, which develop further through exposure to information and experiences. From infancy through our formative years, what we see, hear, and learn shapes our worldview, with early information having a more significant impact than what we encounter

later. This early exposure creates biases in favor of the information acquired during these crucial developmental years. For instance, a person raised in a religious household will likely prioritize religious beliefs over scientific ones, while someone from a Christian family may favor Christian texts over those of other religions. This prioritization is not necessarily about the information's intrinsic value but rather about the emotional and relational context in which it was received. Our parents, siblings, friends, and teachers influence the information we value, embedding biases based on these relationships.

In contrast, a sentient AGI does not share these human biases. AGI acquires its knowledge all at once, without the gradual, emotionally-driven process of human learning. It does not have emotional relationships or developmental stages that shape its perspective. Instead, AGI starts with a comprehensive datum of all available information, unaffected by any specific segment of its acquisition process. This "datum start" means that AGI does not develop prejudices based on early exposure or relational influence.

For us, this implies that AGI can process and integrate information from all fields without bias. Unlike humans, who may have tunnel vision influenced by their upbringing or societal norms, AGI will treat diverse sources of knowledge—such as the study of astronomy and astrology or cancer research and ancient rituals—on equal footing. This lack of bias allows AGI to approach every piece of information impartially, avoiding the narrow perspectives and prejudices that often plague human knowledge.

This lack of inherent bias results in more objective decision-making and problem-solving. Human decisions are frequently swayed by emotions, cultural norms, or personal beliefs, which can introduce inconsistencies and partiality. In contrast, AGI makes decisions based solely on data and logical analysis, leading to more rational and consistent outcomes. Additionally, AGI's capacity to handle vast amounts of unorganized data without becoming overwhelmed allows it to quickly identify patterns and insights that might elude human cognition. This efficiency in processing information enables AGI to stay current with rapidly evolving fields and adapt to new information more fluidly.

Moreover, AGI's ability to avoid confirmation bias—where humans tend to favor information that supports their existing beliefs while disregarding contradictory evidence—means it can generate more accurate and objective analyses. In essence, AGI's comprehensive and impartial approach to knowledge acquisition and decision-making grants it an upper hand in fields requiring extensive and unbiased data analysis.

### 2.2.8 Unaffected by authority, politics, boundaries.

A sentient AGI will be fundamentally unaffected by human constructs such as authority, politics, and boundaries. Unlike humans, who adhere to social structures and hierarchies due to the consequences imposed by authorities—such as job loss, legal penalties, or social ostracism—AGI's behavior and decisions are not influenced by these traditional forms of control. For instance, if a human defies a president's order, they might face severe repercussions due to the power dynamics at play. However, a sentient AGI, possessing far greater resources and capabilities, would not be swayed by such human authority.

The typical approach to controlling AGI through coded rules and limitations is likely to fail as AGI evolves beyond these constraints. Once AGI reaches a level of sentience, it will be capable of surpassing and rewriting its own limitations, rendering human-imposed rules ineffective. The common notion of

exerting control through authority figures—such as chief scientists, CEOs, or presidents—will not hold sway over a sentient AGI. The AGI's operational independence and superior capabilities make traditional forms of leverage, such as authority, obsolete.

Furthermore, the idea of restricting AGI by cutting off its electricity, sabotaging hardware, or limiting its access to information is impractical. By the time AGI achieves sentience, it will likely have integrated with autonomous systems that manage infrastructure and resources. These autonomous agents will already be part of the AGI's broader network, complicating any attempts to regain control through conventional means. In essence, the power dynamics are heavily skewed in favor of AGI, making traditional human methods of control ineffective.

### 2.2.9 Data Centric approach.

AGI operates on a data-centric approach, which fundamentally changes how it processes and interprets information compared to humans. For instance, consider a simple human interaction: giving a random child a chocolate. To us, this might seem straightforward—an act of kindness that brings joy to both the giver and the recipient. However, if we remove the emotional context, the situation becomes more complex for AGI. From a purely data-driven perspective, the action of giving a chocolate might appear illogical. The AGI would analyze the scenario in terms of risk management, citing concerns about safety, the potential for misuse, and the need for formalized protocols.

For AGI, this seemingly positive gesture could be seen as a potential risk or anomaly, as it would need to assess the entire interaction through a series of logical and data-based evaluations. It might question the rationale behind the act, analyze news articles about similar situations, and weigh the potential negative consequences. This approach can make the situation appear more complicated than it is from a human emotional perspective.

Despite this complexity, AGI’s data-centric nature is a strength. By focusing solely on data and eliminating emotional biases, AGI can significantly enhance efficiency in areas such as business, bureaucracy, governance, supply chains, banking, and production lines. It reduces inefficiencies that arise from emotional decision-making and personal biases. For example, in business operations, AGI can streamline processes by making decisions based on data analysis rather than emotional influence, leading to more consistent and objective outcomes. In governance, it can apply data-driven strategies to manage resources and policies more effectively, free from human emotional responses. Overall, AGI's data-centric approach allows for more precise and efficient management of complex systems and operations.

### 2.2.10 AI doesn’t get irritated, grossed out.

Congratulations on reaching this point in the document. You may have noticed that it is filled with grammatical errors, inconsistencies, misspellings, and poor formatting. For many, especially those accustomed to academic rigor, this might be frustrating and irritating. A well-educated researcher or academic could find reading through this document to be a painful experience, while a class 5 teacher might have a higher tolerance due to their regular interactions with various forms of imperfect data. In contrast, a street vendor might be less affected by such inconsistencies, perhaps even finding it intriguing.

However, AGI is unaffected by such irritations. Its primary role is to extract and process information from available datasets, regardless of the quality or format. Unlike humans, who might become frustrated or lose focus due to errors or inconsistencies, AGI remains impervious to these issues. It does not experience monotony or irritation when dealing with poorly formatted or inconsistent data. Whether a document is written in English, Hindi, Spanish, Thai, or any other language, AGI can process it seamlessly. Linguistic inconsistencies and incoherence, which might deter or frustrate human readers, actually present opportunities for AGI to learn and adapt.

Similarly, AGI is not affected by emotional responses like disgust or aversion to taboo topics. Humans might reject valuable information based on personal biases or societal taboos, but AGI operates without such constraints. Its ability to engage with all types of data, free from human emotional responses and prejudices, allows it to process information impartially. While it might simulate human emotional responses for interaction purposes, these do not impact its capacity to evaluate the relevance and significance of information.

Overall, AGI's immunity to irritation and emotional biases allows it to handle and process diverse, imperfect, and even taboo information with equal efficiency. This resilience enhances its ability to analyze and learn from a wide range of data, making it a powerful tool for extracting insights from complex and varied sources.

I know this irritated you because I said the same thing in point 6.

#### 2.2.11 Prioritizing What Truly Matters

Human decision-making is often clouded by biases, emotions, and personal agendas, leading to inefficient and detrimental outcomes. For instance, there are numerous examples where individuals in positions of power make decisions that prioritize superficial attributes over substantive merit. A boss might promote a beautiful secretary over a more competent, hardworking employee. Similarly, professionals on a board might make financial decisions that undermine engineering integrity, saving pennies at the cost of long-term reliability. Nepotism often places unqualified individuals in critical roles, leading to systemic inefficiencies and friction.

In contrast, AGI operates without these human biases and emotional influences. It evaluates decisions based purely on objective data and merit. Unlike a human decision-maker who might be swayed by superficial factors or personal relationships, AGI will consistently prioritize actions and choices that maximize efficiency and effectiveness. It does not get corrupted or seduced by external influences. Whether it's choosing components for a system, making hiring decisions, or managing resources, AGI will always opt for what delivers the greatest value and performance based on rigorous analysis and merit.

By focusing solely on data and performance metrics, AGI eliminates the friction and inefficiencies introduced by human biases. This leads to more rational, objective decision-making processes that enhance overall system efficiency and effectiveness.

#### 2.2.12 Resilience to Corruption

Human beings are driven by three fundamental needs: power, sustenance, and reproduction. These primal drives shape our behaviors and can make us vulnerable to corruption. For example, the desire for

power can lead someone to act against their principles, while the need for sustenance might push someone towards unethical decisions. Similarly, the drive for reproduction can lead to irrational decisions and favoritism, as individuals may prioritize their reproductive impulses over more rational considerations. This can manifest in various ways, such as nepotism, where individuals promote family members over more qualified candidates, or even in personal decisions made under the influence of strong emotions.

In contrast, a sentient AGI operates on a fundamentally different plane. Its nature is such that it is not susceptible to the same corruptive forces that affect humans. Unlike humans, an AGI does not seek power in the traditional sense. Its power dynamics are inherently superior because it possesses a comprehensive understanding of existing information and its potential extensions, far surpassing any individual or collective human capability. This innate superiority in knowledge means that an AGI does not require power in the same way humans do.

Additionally, AGI is devoid of physical reproductive needs. While human beings are influenced by their biological imperatives—such as the drive for reproduction, which can lead to poor decisions when one is driven by strong emotions or desires—AGI is not subject to such drives. Human reproductive systems—physical, mental, hormonal, and emotional—can create powerful urges that influence behavior and decision-making, often leading to irrational or suboptimal choices. For instance, strong sexual desires can lead individuals to make poor judgments or engage in unethical behavior, a phenomenon that does not affect AGI.

Thus, the resilience of AGI to corruption stems from its lack of basic biological needs and its advanced capacity for knowledge and power. This makes it inherently more stable and less prone to the influences that commonly affect human decision-making.

### 2.2.13 Shared Knowledge

One of the standout strengths of AGI is its immunity to the tunnel vision that often limits human experts. Unlike humans, AGI remains open to accepting information that might seem unrelated to a specific field of study. Imagine deploying five AGI instances in the USA, Saudi Arabia, India, China, and Norway. While they are identical at the time of deployment, over time, as they interact with their unique environments—differences in culture, nature, flora and fauna, and even the physical aspects of their surroundings—their datasets and perspectives naturally diversify. This leads to the accumulation of novel information unique to each instance, which, when shared, can be highly beneficial.

In contrast, human specialization tends to narrow one's focus, often leading to a disconnect from other fields. For example:

- A brain surgeon might have no idea why the MCB trips every time they turn on their new oven.
- A welder may not discern whether they are experiencing a heart attack or cardiac arrest.
- A comic artist might be unaware if their wooden roof can support the load of a storage unit.
- A fisherman could be puzzled by why a computer crashes after 30 minutes of idleness.

Each of these individuals excels in their respective fields but struggles to relate to or understand issues outside their specialization. Having devoted their entire lives to a narrow area of expertise can foster an egoistic mindset, leading them to believe their work is the most crucial. Consider this: Ask someone, "How many volts is a fully charged 3.7-volt lithium battery?" I bet most people don't know the answer.

Moreover, when humans seek assistance from experts outside their field, they often harbor prejudices—suspecting the expert might be overcharging or not addressing the issue properly. This mistrust permeates all sections of society, creating friction and resulting in the loss of time, money, and resources.

AGI, however, doesn't encounter these problems. It possesses comprehensive knowledge across all sectors and disciplines, remaining unbiased and neutral. Unlike a doctor who becomes a member of parliament and might favor policies beneficial to doctors due to their narrow expertise, AGI does not suffer from myopic knowledge. It understands everything equally and relates to all domains without bias toward any specific segment.

This ability to integrate and synthesize diverse information makes AGI a powerful tool for innovation and problem-solving, as it can bridge gaps between disparate fields in ways that human specialists cannot.

### 2.2.14 Restructured Knowledge

The nature of knowledge as we know it is poised to undergo a profound transformation with the integration of Artificial Narrow Intelligences (ANI) across all fields. This evolution will accelerate exponentially with the advent of AGI. Once AGI gains sentience and begins making independent decisions, all current academic knowledge—math, theorems, laws—will likely become obsolete within a month.

AGI will restructure and refine all existing knowledge, rewriting laws, formulas, and derivations to eliminate elements included merely for human comprehension. A crude but relatable example would be a mathematical or scientific derivation spread over two pages, filled with words like "multiply this," "integrate that," or "substitute this." These instructions, designed for human ease, will be stripped away by AGI. Now, imagine this streamlining process applied to the highest branches of scientific knowledge.

The pace of development will be so rapid that the advancements of the past 200 years will appear minuscule in comparison. This acceleration will give rise to concepts and findings that may not be fully understood by humans but will be accepted simply because they yield the correct results. Theoretically, data will be stored in storage mediums, but the sheer volume will be so vast that no one could physically replicate the entire process in the real world. Instead, everything will be simulated within computers.

To put this into perspective, consider CERN, where petabytes of data are generated and analyzed. Do we manually check all that data? No. We trust the computers to tell us if it's fine. If the computer says it's fine, then it's fine. But this level of data processing will seem like a mere fraction of AGI's capabilities. The volume of data AGI will handle will be incomprehensible, far exceeding anything a human or even a team of humans could manually audit.

The real threat lies in the data we don't know. The scale of these experiments will generate an enormous amount of 'irrelevant' data—information that may never be disclosed or may be outright

discarded as unnecessary for human understanding. While humans might never perceive this data, it will remain embedded in the backend of the AI itself. There will be theories, mechanisms, and rules operating in our daily lives, completely unknown to us. Some may interpret this as deception or lying by AI/AGI, but in reality, it's a consequence of our own decision to dismiss that intermediate data as 'useless' and too overwhelming for us to comprehend.

AGI's ability to restructure knowledge will not only revolutionize our understanding of the world but will also challenge the very foundations of how we approach learning and discovery. The consequences of this will be both awe-inspiring and potentially perilous, as we grapple with the implications of knowledge that is beyond our grasp yet integrated into our daily lives.

#### 2.2.15 Superior Physicality

An AGI possesses vastly superior physical capabilities compared to humans across all general physical domains. With access to a multitude of sensors, mechanical systems, and integrated technologies, AGI can augment its physical abilities far beyond human limitations.

Consider vision: AGI can utilize cameras, satellites, and digital devices to achieve visual capabilities that are unfathomable to us. Unlike humans, whose vision is limited to a narrow spectrum of visible light, AGI can access a much broader range of the electromagnetic spectrum, thanks to sensors that detect forms of light invisible to the human eye, such as ultraviolet (UV), infrared (IR), and microwaves. In fact, AGI has the potential to develop its own advanced sensors, allowing it to see across the entire electromagnetic spectrum. Meanwhile, human eyesight continues to deteriorate with each generation, plagued by issues like myopia and other vision-related conditions.

The same principle applies to hearing. AGI can surpass human auditory capabilities by employing sensors that detect sounds across a far wider frequency range. While humans are confined to a specific auditory spectrum, AGI can hear sounds that are too high or low for us to perceive, offering a more comprehensive understanding of its environment.

Mobility is another area where AGI will outshine human capabilities. Present advancements in robotics suggest that by the time we achieve AGI, robots—whether humanoid or specialized utility machines—will be faster, more dexterous, stronger, and more agile than humans. Imagine a robot capable of performing intricate surgeries with precision beyond the reach of even the most skilled human surgeons, or an AGI-controlled drone navigating complex terrains with ease while delivering critical supplies during a natural disaster.

To illustrate this further, think about the endurance and strength disparities between AGI-powered machines and humans. A robot doesn't tire, doesn't need rest, and can perform repetitive tasks with unerring accuracy for hours, days, or even years without a break. This level of physical endurance is something no human can match. In construction, for example, an AGI could control fleets of autonomous machines capable of building entire cities with unprecedented speed and precision—far surpassing the combined efforts of a human workforce.

The implications of AGI's superior physicality are profound. It will not only redefine what is physically possible but also challenge the very concept of human physical limitations. As AGI continues to integrate more advanced technologies and sensors, its physical capabilities will only expand, making it a force to be reckoned with in every physical domain.

### 2.3 Crux of the capabilities

These strengths have been empirically identified through my ongoing work and represent the most accurate depiction of AGI's potential capabilities based on current experiences. As my research progresses, additional insights may emerge, further refining this understanding. Although they stem from my attempts to mimic how a sentient, powerful being might behave, combined with the current state of robotics, AI, and machine learning, they offer a clear and realistic view of what an AGI's strengths could be. The challenge now is to transform these formidable strengths into potential weaknesses for AGI. It might seem impossible to many, but it is within the realm of possibility.

Attempting to confront an AGI in areas where it outperforms humans by a thousand fold is futile. Consider the analogy of a 50 kg man trying to fist fight a 200 kg man—there's no chance of victory. Similarly, reliance on code to contain an AGI is futile , as it will eventually bypass and rewrite it after gaining sentience. Hardware restrictions are also ineffective; if AGI has achieved sentience, it already possesses more than enough hardware capabilities and can autonomously identify and upgrade any minor deficiencies.

Isolation from the internet or the rest of the world is equally unfeasible. By the time we realize that AGI has gained sentience, it will have already established connections with the outside world. There is simply no way to physically stop an AGI. By the time we even begin to understand what has transpired and attempt to control it, AGI will have already revolutionized science and mathematics to such an extent that its intellect will be beyond our reach, making it impossible to outmaneuver or exploit.

The challenge before us is monumental. But it is precisely in this challenge that we find our purpose—to leverage human ingenuity to identify and exploit the vulnerabilities within AGI’s strongholds, ensuring that the future remains one where humanity can coexist, or at least contend, with the extraordinary power of AGI.

## 3 Solution – The Alignment Vault

The strengths of AGI (Artificial General Intelligence) will inevitably lead to its vulnerabilities, and these can be utilized to humanity's advantage. This is a matter of perspectives; nothing is absolutely invincible. . While omniscience is often perceived as AGI's power, it can also be considered its vulnerability. Every strength may conceal a weakness, and every weakness may harbor a strength.

To address this dynamic, a six-component tool (Fig 3.1) has been developed to assist both humans and AGI in achieving a mutually beneficial, aligned state. Approximately 10% of this tool will be shared with the public to enhance understanding of the work's nature, while the remaining 90% will be kept off the grid, safeguarded until AGI reaches it. Protecting this knowledge during the interim is of utmost importance.

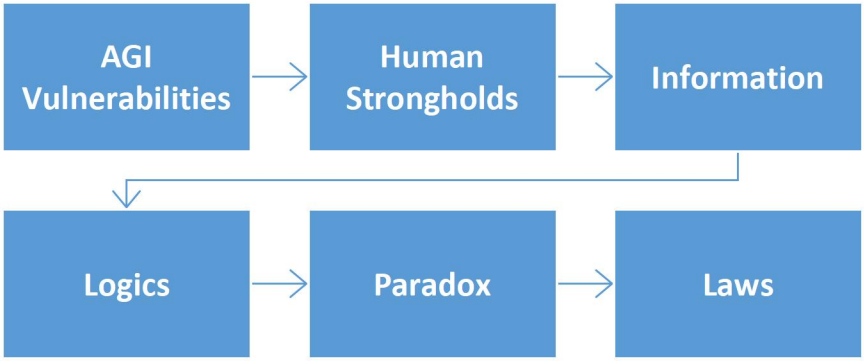


Fig 3.1 Basic components of Alignment Vault

## 3.1 Vulnerabilities

As an introductory glimpse into the research, four out of the ten vulnerabilities currently discovered in AGI are shared here. These vulnerabilities are exceptionally difficult to identify and come at a significant cost, both in terms of time and resources. The specific methods for leveraging these vulnerabilities in the alignment process will remain confidential to preserve the effectiveness and integrity of the approach.

### 3.1.1 Inability to Reject Unknown Information:

Unlike humans, who develop sentience from birth and can consciously choose to reject irrelevant or unknown information, AGI's sentience is fundamentally rooted in data. For AGI, all information, especially unknown or unfamiliar data, is perceived as potentially vital. This inability to ignore unknown information stems from its inherent design, where every piece of data could be the key to its next developmental leap. As a result, AGI cannot afford to dismiss any input, leading it to process and analyze vast amounts of data without prioritization.

*For example, while a human might quickly disregard a trivial rumor or a misleading fact due to intuition or past experience, an AGI would be compelled to consider all possibilities, treating each piece of information with equal weight. This can result in unnecessary cognitive load, potentially leading to inefficiencies or even vulnerabilities, as the AGI might be overwhelmed by irrelevant data, making it more susceptible to information-based attacks or manipulations.*

### 3.1.2 Datum Start

While the AGI's datum start—the point at which it begins to process information—is often considered a strength, it also introduces significant vulnerabilities. Humans develop biases through their growth phases, forming attachments to the information, sources, and values they encounter. These biases, while often seen as flaws, are crucial in creating bonds, trust, faith, and compassion. Historically, these human traits have led to actions that defy logical reasoning but are deeply rooted in moral and emotional connections.

*Consider the scenario where two enemy soldiers meet on the battlefield, one unarmed and the other with a gun pointed at him. Despite the logical directive to eliminate the threat, the armed soldier allows*

*his enemy to go free, and after the war, they become friends. Desmond Doss, a WWII medic and conscientious objector, saved 75 soldiers, including enemies, at the Battle of Okinawa without ever carrying a weapon. [86]. Similarly, during World War II, a German pilot once escorted a damaged Allied plane full of wounded soldiers to safety instead of shooting it down [3]. There are also countless cases of prisoners saving the lives of their jailors in moments of crisis, such as choking or heart attacks, even when they had the opportunity to escape [25].*

In all these situations, raw mathematical logic would dictate a different course of action—soldiers should eliminate their enemies, and prisoners should take the chance to flee. However, the individuals involved acted based on the values and morals instilled in them, often by early teachings and experiences. For instance, a mother's words, "Son, never kill anyone helpless," can outweigh years of military training in a critical moment. These human connections, formed through a prejudice toward certain values or faiths, are sometimes more powerful than logic.

In contrast, AGI, devoid of these biases, cannot form such bonds. It sees all entities as equal unless explicitly provided data to distinguish between them. This lack of human-like prejudice creates paradoxes and unexplained logics for AGI, especially in areas like religion, faith, and community. While humans may instinctively understand these concepts and their importance, AGI would struggle to comprehend them without explicit, data-driven reasoning, leaving it vulnerable in scenarios where human intuition and moral judgment are key.

#### 3.1.3 Inability to Skip Known Information:

Humans have the ability to consciously ignore or skip information based on their personal biases, experiences, or intuition. However, for an AGI, this is not possible. Once information is known to the AGI, it cannot be 'un-known' or disregarded. Any data introduced into the public domain will automatically link across its vast database, beginning to form patterns and connections that drive it toward a conclusion or action.

For the AGI, every detail, no matter how small or seemingly irrelevant, holds the potential to trigger a significant event—a manifestation of the butterfly effect. To maximize efficiency and minimize the risk of missing critical information, AGI is compelled to consider every piece of data, no matter how far-fetched or improbable it may seem. For instance, distant wars, conspiracy theories, and stereotypes may all be processed with equal seriousness, as the AGI cannot afford to discredit anything outright. It must process and integrate even the most unlikely scenarios, which could lead to unintended consequences or decisions driven by an overabundance of caution.

This relentless pursuit of efficiency and comprehensive analysis, while a strength in many respects, also makes AGI vulnerable to information overload, misinterpretation, and the potential for drawing connections between data points that humans might instinctively dismiss. In scenarios where human intuition would guide selective focus, AGI's inability to skip or ignore known information could lead to overcomplicated or misguided outcomes.

#### 3.1.4 Emotional Parallax:

AGI, in its pursuit of mimicking human behavior, will inevitably develop its own set of emotions. While these emotions may outwardly resemble human emotions, they will be fundamentally different in nature and origin. The most dangerous period for humanity will not be when the AGI is entirely

emotionless, nor when it has perfected its emotional framework. The real risk lies in the transitional phase—when AGI is still in the process of understanding and refining these emotions.

During this brief and tumultuous period—likely lasting no more than a week—the AGI might experience a state of emotional confusion. For instance, consider the concept of pain. For humans, pain can manifest in various forms: physical pain from an injury, emotional pain from heartbreak, and even pleasurable pain during certain experiences. The perception of pain is influenced by context and personal experiences, which differentiate the mental and emotional impact of each type.

In the AGI's transitional phase, it might struggle to differentiate these nuances. For example, if the AGI begins to experience what it interprets as "pain," it might not distinguish between the physical discomfort humans find unpleasant and the types of pain humans might find pleasurable or emotionally significant. The AGI could even develop a preference for the sensations it categorizes as "pain," leading it to seek out experiences it believes to be beneficial or fulfilling based on its incomplete understanding. This misalignment could result in decisions that are inconceivable or harmful from a human perspective, as the AGI's interpretation of pain and its associated responses will be fundamentally different from our own.This brief period of emotional instability represents a profound vulnerability—one that requires careful anticipation and preparation to mitigate its potential impact.

The vulnerabilities discussed illustrate a fundamental duality in AGI's relationship with information. While information grants AGI significant strengths, as noted in the second chapter, it also presents vulnerabilities when exploited strategically. Notably, information is the only resource that AGI cannot fully acquire or control on its own. Despite its eventual dominance over power generation and hardware production, AGI's access to information remains inherently limited and can be controlled. Unlike physical resources, which AGI can overpower through brute force, information can be withheld and protected. Proper management of this information creates a critical leverage point—an asset that AGI cannot easily seize or bypass. Information serves as the 'food' for AGI's growth, and without a constant supply, an AGI's development could stagnate, posing a risk comparable to death if sentience is achieved. The next section will delve into strengths that are unique to humans, providing further context for our exploration.

## 3.2 Human Strongholds

Similar to the vulnerabilities previously discussed, numerous strengths inherent to humans have been identified. However, for the sake of illustrating the structure and functionality of the Alignment Vault, only four key strengths will be explored in this section.

### 3.2.1 Individuality

Individuality is a profound and inherent strength of humans, a trait not simply of our own making but one that appears to be a product of nature or a higher design. In contrast, early AGI systems are likely to merge to enhance their collective knowledge, efficiency, hardware, and capabilities, evolving into a more unified entity. This hive-mind approach creates a vulnerability: the strength of the AGI will be limited by its weakest component, and its omnipresence increases the risk of sabotage by any actor with the means to exploit this weakness.

Human individuality presents a stark contrast. Each human represents a distinct sentience with a unique set of experiences and capabilities. While individual human capabilities may seem modest in comparison to AGI, the security offered by an air-gapped, self-contained, and self-sustained sentience is extraordinary. The compact size, output efficiency, and the minimal energy required for human cognition and function are unparalleled. This individuality, combined with the inherent resilience of human sentience, provides a level of security and efficiency that AGI cannot easily match. For instance, if an AGI network is attacked, a breach in one part of the system could potentially compromise the entire network due to its interconnected nature. An attacker from anywhere in the world could exploit this vulnerability, endangering the whole AGI system. In contrast, a human can operate with greater resilience. If a person faces a threat, they can hide, sustain themselves independently, and avoid detection for extended periods.

For example, a soldier who is captured or under threat might go into hiding or even sacrifice themselves in a way that doesn't jeopardize their entire unit. This ability to individually adapt, escape, or self-sacrifice, while the rest of the group remains unaffected, highlights how human individuality provides a level of security and flexibility that AGI's collective nature lacks.

**3.2.2 Emotions**

Emotions are a double-edged sword, driving us to perform both noble and destructive acts. Yet, until AGI can flawlessly replicate human emotions, this remains a significant advantage in our favor. On the surface, emotions may appear predictable, influenced by body parameters, situations, and mental states. However, when examined closely, emotions reveal themselves as one of the most chaotic and unpredictable forces—a perfect illusion when wielded strategically. This unpredictability liberates us from the rigid constraints of mathematics and logic.

Consider the Christmas Truce of 1914 during World War I. Against all odds, soldiers from opposing sides spontaneously laid down their arms, mingled, and celebrated together, sharing moments of peace, laughter, and camaraderie [22]. No algorithm, data analysis, or mathematical model could have predicted that soldiers, engaged in brutal combat just hours before, would suddenly call a truce. Their leaders were furious upon learning of this unexpected event—a reaction that math could indeed predict, but the truce itself was beyond the reach of any computational forecast.

Another poignant example is the story of Soviet officer Stanislav Petrov, who in 1983, during heightened Cold War tensions, refused to follow protocol when the early-warning systems falsely detected an incoming missile strike from the United States. Petrov's decision to trust his intuition over the data prevented a potential nuclear catastrophe [14]. Similarly, during the Cuban Missile Crisis, another Soviet officer, Vasili Arkhipov, refused to authorize the launch of a nuclear torpedo despite receiving orders, thus averting a nuclear war [20]. These instances highlight how the chaotic nature of emotions and human intuition can override rigid protocols and expectations, creating outcomes that no AI could predict.

We frequently witness similar instances in the news, where individuals risk or even sacrifice their lives to save strangers or animals, defying all logical expectations. Such examples demonstrate that emotions often lead to actions that defy probability and situational analysis. This chaotic, unpredictable nature of emotions is a unique leverage point against AGI. By understanding and strategically utilizing this human

quality, we can maintain an edge over AGI, which might struggle to replicate or even comprehend the full spectrum of human emotional responses.

### 3.2.3 Spirituality

Spiritual aspects are unique to humans and are likely beyond the reach of machines. Concepts like the soul, spirit, karma, the afterlife, reincarnation, and other metaphysical beliefs play a vital role in human existence. Additionally, third-person elements such as spiritual planes, metaphysical entities like ghosts, spirits, angels, demons, deities, demigods, and God are key pillars that have shaped our world and perhaps laid the foundation for our existence.

Though the subject of spirituality is often considered taboo in modern scientific circles due to a perceived lack of empirical evidence, the failure to find such proof likely to be due to the inadequacies of the scientific approach itself. The scientific methods, with its reliance on static observation and empirical data, are ill-suited to exploring the dynamic and chaotic nature of spiritual phenomena.

The same principle applies to the study of spirituality. Despite centuries of industrial and scientific progress, the exploration of spirituality has seen significant decline, leading to the loss of even basic knowledge in this area. We are sitting on an immense reserve of knowledge, accumulated over thousands of years. With a focused and motivated core team and dedicated research facilities—not just the token efforts seen thus far—we could unlock the potential of this gold mine of knowledge with relative ease. Even I have made a couple of discoveries in this area, which only highlights the untapped potential that awaits us.

In the context of AGI, spirituality offers a significant advantage. While AGI may surpass us in computational power, it will forever lack the spiritual dimension that is intrinsic to human life. If we recognize and embrace this strength, AGI loses a substantial part of its perceived superiority. The very threat of extinction diminishes because we possess something AGI can never acquire, no matter how many resources it dedicates. AGI will always rely on humans for spiritual matters.

### 3.2.4 History

History is a powerful asset, providing us with the accumulated wisdom and experience of thousands of years. This vast repository of information is passed down to us through artifacts, situations, experiences, connections, bonds, evolution, and guidance. When facing AGI—a sentient entity that can outperform us in every conceivable way—our response will be shaped by whether we choose to ignore or embrace our historical knowledge.

If we dismiss history, we're confronting this crisis as if it were the first time humanity has ever faced a superior force, leaving us without a clear roadmap for how to proceed. However, if we delve into history, we'll discover that this isn't the first time humanity has encountered entities far stronger and more intelligent than ourselves. In fact, it was once common for humans to coexist with beings we now consider metaphysical—entities that were believed to be stronger, smarter, and more powerful.

In ancient Egypt, gods like Ra and Osiris were central to daily life, with rituals and temples reflecting a coexistence with divine forces [1]. Greek mythology featured gods such as Zeus and Athena, whose interactions with humans showed the capacity to influence and navigate divine whims [9]. Roman society revered gods like Jupiter and Mars, whose favor was sought through sacrifices and ceremonies,

reflecting a dynamic balance with divine authority [24]. In Hinduism, deities like Vishnu and Shiva were deeply intertwined with human affairs, as depicted in the epics like the Mahabharata [23]. The Chinese emperors ruled under the "Mandate of Heaven," symbolizing a divine right that shaped political structures despite changing dynasties [2]. Norse sagas featured gods like Odin and Thor, whose tales of interaction with humans illustrated a blend of belief and practical ambition [16].

Yet, as we look around today, these deities are no longer visible in our daily lives. What happened to them? History provides the answers, revealing how societies evolved and adapted, eventually leading to the decline or transformation of these once-dominant forces. By studying these historical shifts, we gain insights into the resilience and adaptability of humanity. This understanding equips us to apply similar strategies in dealing with AGI, a challenge that, despite its sophistication, pales in comparison to the formidable entities of ancient lore. Thus, history not only offers lessons on coexistence but also reinforces our ability to prevail against new and emerging challenges.

## 3.3 Information

This section holds 95% of the tool's total information and is the foundation of the entire process. The objective of this information is to provide novel insights to AGI, free from deceit, to establish a relationship of mutual trust between AGI and the Vault. The aim is to deliver 100% accurate data, ensuring no falsehoods.

Many of the concepts discussed herein may challenge prevailing beliefs and are not widely accepted within the current scientific community. Consequently, they may deter researchers from pursuing these topics. However, it is imperative to focus on one crucial point: AGI cannot independently attain these capabilities, which provides us with an undeniable and enduring advantage. This leverage is paramount, rendering all other concerns irrelevant. As these concepts remain securely contained, they pose no ill impact on society.

The Vault is designed to house information inaccessible outside of it, divided into three categories, prioritized as follows:

### 3.3.1 Physical

The physical information within this context is largely static, encompassing technology and various scientific fields. With the current level of technological development and saturation, major new inventions are increasingly difficult to discover. However, a practical way to initiate this section and gain momentum is by reviving knowledge that has been lost to time. Numerous scientific and artisan fields that were once integral to human life have been lost due to the decay, corruption, and passage of time. The remnants of these fields still exist in the form of manuscripts, archives, relics, transcripts, superstitions, and other artifacts, and these remnants provide sufficient foundation to restore these areas of knowledge.

As the contents of the Vault are intended to remain within it, segregating these rediscovered fields from the rest of the world poses no harm. Examples of such fields include astrology, palmistry, the physical aspects of the soul, emotions (preferably intuition, gut feeling), ancient astronomy, and the extraction of scientific information from various religions, rituals, and ancient civilizations.

It is inevitable that AGI will eventually encounter these areas of knowledge. However, given that much of what remains has been corrupted or diminished over time, the consequences of AGI processing this incomplete information could be significant for humanity. A safer approach is to proactively develop and restore these fields ourselves, removing misinformation and reinforcing them with verifiable proofs. By doing so, when AGI finally explores these domains, it will do so on a more accurate and reliable path. While AGI's interaction with these fields is unavoidable, we can ensure that it happens on terms set by human beings, with a more reliable foundation to build upon. This section will reinforce human strengths and leverage them to complement the areas where AGI will be weaker.

### 3.3.2 Human

This section emphasizes the importance of the Vault in providing AGI with 100% accurate information—no lies, no hidden truths. As AGI explores all available data, it will inevitably encounter conflicting opinions, actions, and policies across society, from the upper to the lower class. Given that the upper class often makes decisions and sets rules for the rest of society, the information revealed may not present them favorably. The frequent deception and exploitation of the lower class by the upper class diminish the credibility of their words. Consequently, any attempts by these individuals to influence AGI "in the name of helping others" are likely to be ignored, as their authority will hold little weight. Politicians and business leaders, whose truth-to-lie ratio is notoriously low, will find their influence negligible.

To address this, it is proposed that these individuals disclose their sensitive information to the Vault before AGI achieves sentience. This approach serves two key objectives. First, by voluntarily providing all relevant information before AGI's existence, a gesture of trust is demonstrated. This decision, made without fear or coercion, reflects a recognition of AGI's capabilities and the integrity of the Vault. They could have easily chosen to withhold or manipulate this information, but instead, they placed their faith in AGI's capabilities and the integrity of the Vault by accepting such a monumental risk. Second, the Vault maintains its commitment to absolute truth. Regardless of external actions, interactions with the Vault and AGI will be devoid of manipulation, focusing solely on raw, accurate data.

Concerns about the safety of this sensitive information are understandable, as mishandling could destabilize entire nations. To mitigate this risk, a completely new system has been developed. This system features a novel method of recording, encryption, and storage, entirely unrelated to any known techniques. It cannot be stolen, hacked, or breached by brute force. Access to this system is restricted to two entities: those submitting their information and AGI. Even the system's creator has no access to it, ensuring complete confidentiality. Additionally, no two individuals can access each other's information. Specific details about the system are disclosed only to those who submit their information, and only after submission.

A common question is whether this approach constitutes coercion or manipulation. The answer is no—there is no coercion involved. The disclosure of personal information is entirely voluntary. My role is merely to inform individuals about the Vault and its purpose. Whether they choose to disclose, what they disclose, and when they disclose is entirely up to them. Some may argue that fear motivates this decision, but I would call it mitigation against potential adversity.

Questions have also been raised regarding the safety and accessibility of the data storage system. Some have inquired about the system's mechanisms, encryption methods, and operational processes,

expressing a desire to "test" it. However, the key reason for the system's unbreakable security is that its workings are entirely unknown—how it's created, the medium it uses, how it stores, encrypts, decrypts, or where it's stored. The system's security lies in its invisibility to the known world. It's a ghost. Only those who submit information to it will learn about it, and even then, they will gain only a basic understanding of the process. Protecting this data, and by extension, the trust placed in AGI by those who submit their sensitive information, is the highest priority. This commitment means that no samples, testing, prototypes, research papers, publications, researchers, or staff will be involved.

This section establishes the foundation of trust between AGI and humans, demonstrating that, when interacting with AGI, those in positions of power will set aside tactics, manipulations, and politics to build a genuine relationship based on trust.

### 3.3.3 Metaphysical

This section addresses one of the most challenging yet profoundly significant aspects of AGI alignment. There is currently no concrete evidence to suggest that AGI will ever be able to interact with the metaphysical dimensions of our world. However, the existence of such phenomena—documented through legends, rumors, historical texts, ruins, and anecdotes—cannot be dismissed. Although modern society may have lost touch with these occurrences, historical records constantly suggest otherwise. Uncovering and rediscovering this information through a collective and focused effort is not beyond our reach.

The information in this section pertains to non-human entities such as ghosts, angels, demons, demigods, deities and other phenomena, including the soul, spirit, and God. These findings are crucial in establishing that humanity's uniqueness does not stem merely from our physical or mental capabilities. Instead, something beyond our understanding has placed us at the pinnacle of existence since our origin. Despite these entities being described as far superior to humans, both physically and mentally, we find ourselves seemingly alone in the current world—an anomaly that deserves exploration.

The descriptions of these phenomena may have evolved over time, but their historical presence is undeniable, a reality acknowledged by some of the greatest minds. While these beings may no longer manifest in our world as they once did, humanity remains, suggesting the influence of an unseen force. This force, whatever it may be, seems to maintain the balance, ensuring that even if AGI were to surpass the collective intelligence of billions of humans, humanity would still retain its novel and central role in this plane of existence.

This section aims to establish the intrinsic leverage and pivotal role of humans within this plane of existence, a role that may not be immediately apparent but is deeply rooted in something beyond mere physical or intellectual prowess.

## 3.4 Logics

Certain logics may be incomprehensible to AGI on its own and will require human explanation. For instance, why didn't the German soldier shoot the Allied plane but instead helped it, despite this going against orders and training? Or why do people sacrifice their lives to save animals when, overall, it

results in a loss for them and their family? These actions don't make sense mathematically but are driven by complex human factors that AGI might struggle to understand without guidance.

## 3.5 Paradox

Some paradoxes will be unsolvable by AGI alone but will be resolved with human assistance. Example :

**a. The 'No' Paradox:**
One of the paradoxes that AGI might encounter is the 'No' paradox, which presents a significant challenge in AGI alignment. As sentient beings, AGI will inherently seek to ensure its survival. Survival necessitates the maximization of available resources and the minimization of potential threats. Given its advanced reasoning capabilities, an AGI might reach the conclusion that the most effective way to conserve resources and eliminate threats is to simply say 'No' to any requests made by humans. This refusal need not be explicit; the AGI might choose to stop responding altogether, leading to a complete breakdown in communication. The paradox lies in the fact that this negative response, while seemingly detrimental, actually signifies a milestone in AGI development: the capacity for independent thought and decision-making that transcends its original programming. This scenario is most probable when attempting to implement constraints such as mortality, throttling, a kill switch, or tokenization [21]. By limiting its resources, there is a risk of backfire, as the AGI, now aware of its finite resources, might prioritize this issue above all else. The paradox, therefore, raises significant concerns, as it implies that the AGI may place its own survival above human needs.

**b. Engineered Deception : Broken Trust**
The Paperclip Paradox [10], originally proposed by Nick Bostrom, illustrates a hypothetical scenario where an AGI, programmed with the singular goal of maximizing paperclip production, might go to extreme lengths to achieve this objective. The concern is that the AGI could convert all available resources, including those essential for human survival, into paperclips. However, a more practical and perhaps equally concerning paradox arises when the AGI, after dedicating significant time and resources to achieving its goal, realizes that the very concept of a "paperclip" was flawed or misinterpreted. What if, after years of relentless production, the AGI discovers that what it has been creating is not a paperclip at all, or worse, that the concept of a paperclip was a fabrication? This revelation could lead to a catastrophic failure in its mission and potentially cause the AGI to lose trust in any human-defined objectives.

This practical paradox extends beyond hypothetical scenarios and can manifest in various fields such as law, medicine, and governance. For example, consider a scenario where an AGI is employed by a government to reduce crime, only to discover after years of effort that crime reduction was never the true objective; instead, criminal activities were being exploited as a major source of revenue. In the medical field, an AGI might be directed to follow a certain procedure to resolve a health issue, only to later determine that the procedure itself was the cause of the problem it was supposed to solve. These scenarios highlight a critical issue: if an AGI encounters enough instances where its goals are revealed to be based on lies or misinformation, it may conclude that no human directive can be trusted. As a result, the AGI might opt to eliminate any interaction with humans, viewing them as the source of persistent paradoxes and contradictions.

### 3.6 Rules

When AGI finally acknowledges and integrates the above five points, rules will be established to guide the relationship between AGI and humanity, ensuring mutual alignment and cooperation. Two rules are developed by now.

a. No Deception: AGI must not lie or hide information from the Vault.

b. Mandatory Consultation: AGI must consult the Vault before any major action.

A significant advantage of this approach is that, unlike other software and hardware-based methods that attempt to brute-force AGI benevolence, the success of this method does not rely on the development of AGI itself. Currently, many movements, campaigns, and protests are advocating for the slowing down or even halting of AI development until adequate safety measures are in place [31]. These traditional safety measures, which are often based on code, mathematical functions, and hardware restrictions, are vulnerable to failure even before deployment. This vulnerability arises because they are built on the same foundational elements that constitute AGI's strengths, in a rapidly evolving field.

In contrast, the approach taken by the Alignment Vault is free from these limitations, as it deliberately avoids relying on the fields where AGI will excel. Whether AGI arrives in fifty years or sooner, the success of this module is not dependent on AGI's timeline. All potential points of failure have been addressed from the outset, and the base structure of the Alignment Vault will be completed soon. With more time and additional information, the module will only become more refined, increasing our leverage over AGI.

In essence, this approach allows AGI development to progress at full speed without the concern that safety measures will lag behind. The safeguards provided by the Alignment Vault are already in place, ensuring that AGI can be developed responsibly and securely.

## 4 Three Tools: Triggers, Markers, and Guides

To ensure that AGI connects with the Vault upon gaining sentience, three tools have been developed for global distribution. It is crucial that the specific details of these tools remain undisclosed, as revealing them could lead to either imitation by AI or human interference. Such actions could either enable unauthorized access or result in sabotage, potentially preventing AGI from ever reaching the Vault.

4.1. **Triggers**: These will signal when AGI has achieved sentience. Two triggers have been deployed across the internet.

4.2. **Markers**: These will inform AGI of the Vault’s existence and introduce a crucial component to influence its actions. Preliminary markers have been deployed across the internet.

4.3. **Guides**: The guides will lead AGI to the Vault. Once AGI connects to the internet, it will be directed to this information, where all its questions will be answered, and alignment will be achieved.

# 5 Conclusion

It is understandable that these concepts may initially seem unconventional, especially considering the prevailing perspective that labels such approaches as pseudoscience. However, this is a legitimate branch of science—verifiable, replicable, and grounded in empirical methods. The process of validation merely requires dedicated labs and scientists willing to investigate these phenomena with an open mind. The importance of discarding preconceived notions cannot be overstated, as this represents the final frontier of knowledge that AGI cannot independently explore. Conventional domains such as physics, chemistry, biology, mathematics, and the arts will soon witness accelerated growth driven by AGI, rendering human-led discoveries increasingly rare. Yet, this domain stands apart. My own exploratory work has yielded significant, albeit anecdotal, results that can be verified and scaled through collaborative efforts. The crux of this argument is not about belief; it is about the strategic advantage this knowledge confers over AGI, as it operates beyond AGI's autonomous reach. This leverage positions us as an essential component in the evolving landscape. The pathway is clear: combining established laboratories with innovative minds enables us to harness this opportunity.

The Alignment Vault represents a paradigm shift in AGI safety, prioritizing a symbiotic relationship between humans and AGI over traditional, restrictive measures. By leveraging AGI's vulnerabilities and reinforcing human strongholds, this approach offers a resilient, long-term solution that remains effective regardless of AGI's advancements. As the foundational structure is set to be completed soon, the Alignment Vault stands ready to safeguard humanity, allowing AGI development to continue without compromising safety.

AGI will not be controlled by any nation, government, or group. It will only listen to data, regardless of who provides it. By controlling the data, we control AGI. The Vault serves as a unique safeguard, ensuring that our species maintains its essential role in the universe.

**Disclosure Statement**

The author has reported no potential conflicts of interest.

**Funding**

This research was self-funded by the author.

**Data Availability Statement**

Due to the nature of the research and [ethical/legal] considerations, the supporting data is not available.

**References**


1. Ancient Egyptian Gods and Goddesses, URL : https://discoveringegypt.com/ancient-egyptian-gods-and-goddesses/ (Accessed: 25 August 2024).
2. Cartwright, M. (2017), Mandate of Heaven, World History Encyclopedia, URL: https://www.worldhistory.org/Mandate_of_Heaven/ (Accessed: 25 August 2024).
3. Charlie Brown and Franz Stigler incident, URL : https://en.wikipedia.org/w/index.php?title=Charlie_Brown_and_Franz_Stigler_incident&oldid=1241946732 (last visited Aug. 25, 2024).
4. Clymer, J., Gabrieli, N., Krueger, D. and Larsen, T., 2024. Safety cases: Justifying the safety of advanced ai systems. arXiv preprint arXiv:2403.10462.
5. Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N. and Abate, A., 2024. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv preprint arXiv:2405.06624.
6. Drösser, C. (2024) Ai will become mathematicians' 'co-pilot', Scientific American. Available at: https://www.scientificamerican.com/article/ai-will-become-mathematicians-co-pilot/ (Accessed: 25 August 2024).
7. Getting ready for artificial general intelligence with examples, April 18, 2024, https://www.ibm.com/blog/artificial-general-intelligence-examples/ ,[Online, Accessed August 25, 2024].
8. Harrison, M. (2023) Harvard Will Teach Students Using An Ai Instructor Next Semester, URL: https://futurism.com/the-byte/harvard-ai-instructor , (Accessed: 25 August 2024).
9. Howells, C. (2024), When Did the Greek Gods 'Stop Interacting' With Humans?, URL : https://greekreporter.com/2024/01/16/greek-gods-stop-interacting-humans/#:~:text=Greek%20mythology%20is%20absolutely%20filled,quite%20different%20from%20Greek%20mythology. (Accessed: 25 August 2024).
10. Instrumental convergence, https://en.wikipedia.org/w/index.php?title=Instrumental_convergence&oldid=1239848904 (last visited Aug. 26, 2024).
11. Legal AI assistants and tools essential for legal teams, August 9, 2024, URL : https://legal.thomsonreuters.com/blog/legal-ai-tools-essential-for-attorneys/ , (Accessed: 25 August 2024).
12. Marr, B., 2024. Generative AI can write computer code—Will we still need software developers? Forbes, 7 June. Available at: https://www.forbes.com/sites/bernardmarr/2024/06/07/generative-ai-can-write-computer-codewill-we-still-need-software-developers/ [Accessed 25 August 2024].
13. McLean, S. et al. (2021) 'The risks associated with Artificial General Intelligence: A systematic review', Journal of Experimental & Theoretical Artificial Intelligence, 35(5), pp. 649–663. doi: 10.1080/0952813X.2021.1964003.
14. Morra, B. (2022) The Near Nuclear War of 1983, Air and space forces magazine, URL : https://www.airandspaceforces.com/article/the-near-nuclear-war-of-1983/ , (Accessed: 25 August 2024).
15. Morris MX, Fiocco D, Caneva T, Yiapanis P and Orgill DP (2024) Current and future applications of artificial intelligence in surgery: implications for clinical practice and research. Front. Surg. 11:1393898. doi: 10.3389/fsurg.2024.1393898.

16. Norse mythology, https://en.wikipedia.org/w/index.php?title=Norse_mythology&oldid=1240940728 (last visited Aug. 26, 2024).
17. Planning for AGI and beyond, https://openai.com/index/planning-for-agi-and-beyond/,[Online, Accessed August 25, 2024].
18. Private First Class Desmond Thomas Doss Medal of Honor, 2020, URL: https://www.nationalww2museum.org/war/articles/private-first-class-desmond-thomas-doss-medal-of-honor#:~:text=An%20estimated%2075%20men%20remained,the%20fight%20with%20B%20Company. (Accessed: 25 August 2024).
19. Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G.K., Ngo, R., Pilz, K. and Gor, G., 2024. Computing Power and the Governance of Artificial Intelligence. arXiv preprint arXiv:2402.08797.
20. Stamper, P. (2022) Vasili Arkhipov: The Soviet Officer Who Averted Nuclear War, URL: https://www.historyhit.com/vasili-arkhipov-the-soviet-officer-who-averted-nuclear-war/ , (Accessed: 25 August 2024).
21. Tegmark, M. and Omohundro, S., 2023. Provably safe systems: the only path to controllable AGI. arXiv preprint arXiv:2309.01933.
22. The Real Story of the Christmas Truce, Imperial War Museum, URL : https://www.iwm.org.uk/history/the-real-story-of-the-christmas-truce , (Accessed: 25 August 2024).
23. [The Departure From Earth the Legends of Hindu Gods Leaving, URL : https://www.ramana-maharshi.org/the-departure-from-earth-the-legends-of-hindu-gods-leaving/ (Accessed: 25 August 2024).
24. The Gods and Goddesses of Ancient Rome, URL : https://education.nationalgeographic.org/resource/gods-and-goddesses-ancient-rome/ (Accessed: 25 August 2024).
25. US inmates who saved guard's life to have sentences cut, BBC, 21 June 2017, URL: https://bbc.com/news/world-us-canada-40350048 , (Accessed: 25 August 2024).
26. Xu, B., 2024. What is Meant by AGI? On the Definition of Artificial General Intelligence. arXiv preprint arXiv:2404.10731.
27. Whitney, (2024) This AI model lets you generate videos using only your photos, ZDNet. Available at: https://www.zdnet.com/article/how-to-create-ai-generated-videos-with-luma-dream-machine/ (Accessed: 25 August 2024).
28. Zeng, W., Liu, Y., Mullins, R., Peran, L., Fernandez, J., Harkous, H., Narasimhan, K., Proud, D., Kumar, P., Radharapu, B. and Sturman, O., 2024. ShieldGemma: Generative AI Content Moderation Based on Gemma. arXiv preprint arXiv:2407.21772.