SUPERINTELLIGENCE
AN iQ COMPANY

SUPERINTELLIGENCE DESIGN WHITE PAPER #6: CATALYSTS FOR GROWTH OF SUPERINTELLIGENCE

by Dr. Craig A. Kaplan
May 2025

Note: This white paper was released quickly to share our designs and inventions for safe AGI and SuperIntelligence as soon as possible. It has not yet been formatted according to formal journal standards. All references are already included in this document, though the numbering will be updated for consistency in a future version. All figures for White Paper 6 are located at the end of the document but may not be directly referenced in the text. In contrast, White Paper 10 (Planetary Intelligence) includes all figures and descriptions using a different numbering system. We hope this document helps researchers and developers pursue safer, faster, and more profitable approaches to building advanced AI, AGI, and SI systems that reduce p(doom) for all humanity.

TABLE OF CONTENTS

ABSTRACT

Data is the “fuel” that powers the machine learning “engine” for Artificial Intelligence. However, identifying high-quality data that can catalyze smarter AI, AGI, and SuperIntelligent systems is becoming an increasingly challenging bottleneck for machine learning. This whitepaper not only describes novel methods for identifying the most valuable data, but it also presents an entirely new framework for understanding the information content of AI-relevant datasets. The methods can be used by intelligent systems autonomously or in collaboration with humans. Novel methods for accelerating AI learning and updating the knowledge of AI systems in real time are also disclosed. Consistent with the view that human survival may depend on the fastest path to AGI, also being the safest path, the white paper describes catalysts that help maximize alignment between the values of AGI and humans. These innovative catalysts increase not only the intelligence but also the safety of AI systems.

SUMMARY

White Paper #6 describes a novel approach to developing safe and ethical Artificial General Intelligence (AGI) and SuperIntelligent AI systems. It emphasizes the importance of collective intelligence, a network of human and AI agents, instead of relying on a single, monolithic LLM.

White Paper #6 focuses on three key aspects of AI systems:

1. **Information Acquisition:** The white paper proposes new methods for identifying and acquiring relevant and useful information for increasing AI systems' intelligence. This includes expanding the traditional Shannon-sense information theory to incorporate “differences” as a measure of information, rather than simply relying on probabilities and surprise.
2. **Representation:** The white paper highlights the importance of using high-level representations for problem-solving instead of relying solely on low-level representations like bits or tokens. It suggests that adopting such representations can significantly accelerate the development of intelligent systems.
3. **Safety:** The white paper emphasizes the importance of ensuring safety and ethical considerations in designing and developing AI systems. It proposes using a combination of human oversight, automated simulation methods, and adversarial testing techniques to achieve alignment between human and AI values, thus reducing the risks associated with the uncontrolled growth of SuperIntelligent AI.

1.0 OVERVIEW OF THE INVENTION

The current invention focuses on means for increasing the intelligence of Artificial Intelligence (AI), Artificial General Intelligence (AGI), and SuperIntelligent (SI) systems as rapidly, effectively, and SAFELY as possible. Note that AI, AGI, SI, and PSI are used interchangeably in this disclosure since the inventive methods relate to all these forms of Artificial Intelligence.

Since the current invention builds on the work of existing pending patents, I begin by citing those PPAs.

Next, as background for the invention, I disclose a previously unpublished analysis of the theoretical underpinnings of the collective intelligence approach to AGI and SI, including a brief introduction to Information Theory.

Next, a more detailed discussion of Information Theory contrasts Classical Information Theory (e.g., as developed by Claude Shannon) to a novel and inventive approach called Kaplan Information Theory (KIT). Classical Information Theory is described as a subset of the more general KIT approach.

KIT enables novel catalysts for increasing the intelligence of an AI, AGI, or SI system. KIT extends some of the methods from Classical Information Theory in non-obvious ways. KIT also enables entirely new methods for increasing a system's intelligence. These inventive methods are explained, and some preferred implementations are described.

Finally, detailed implementation examples show how intelligent systems (e.g., AI, AGI, and SI) can be customized via the inventive methods. AI, AGI, and SI systems can use the methods to increase their intelligence both autonomously and in collaboration with humans.

2.0 PREVIOUS PPAS (INCORPORATED BY REFERENCE)

The fastest and safest path to the development of Artificial General Intelligence (AGI) and SuperIntelligent AGI (SuperIntelligence or "SI") has been described in previous invention disclosures. Methods for increasing the intelligence of AI systems generally, as well as the development of AGI and Personalized SuperIntelligence (PSI), have also been previously disclosed. Therefore, the following PPAs are incorporated into this PPA by reference.

This provisional patent application (PPA) incorporates by reference all work in the PPA # 63/487,494 entitled: Advanced Autonomous Artificial Intelligence (AAAI) System and Methods, which was filed and received by the USPTO on February 28, 2023.

The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Ethical and Safe Artificial General Intelligence (AGI), Including Scenarios with Technology from

Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, X, Tesla, Nvidia, Tencent, Apple, and Anthropic, which was filed with the USPTO on March 17, 2023.

The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Human-Centered AGI, which was filed with the USPTO on May 24, 2023.

The PPA also incorporates by reference all work in the PPA entitled: System and Methods for Safe, Scalable, Artificial General Intelligence, which was filed with the USPTO on July 18, 2023.

The PPA also incorporates by reference all work in the PPA #63/519,549 entitled: Safe Personalized Super Intelligence (PSI), which was filed with the USPTO on August 14, 2023.

The current PPA contains further inventions that can be used with the system and methods described in the above-mentioned PPAs as well as in a standalone fashion.

3.0 BACKGROUND FOR THE INVENTION

In an as-yet-unpublished analysis, the inventor has described the background of the current invention, including how it draws on seminal ideas from some of the original founders of the field of AI. While the current invention goes far beyond these inspirational notions in novel and useful ways, understanding of the roots of the field and of the general collective intelligence approach to AGI may be helpful. The unpublished text by the inventor in this “Background for the Invention” Section is being submitted to a conference, but has not yet been made public, as of the time that this PPA was filed.

This unpublished analysis frames the contributions of four of the Founders of AI as “gifts” that can be extrapolated in novel, non-obvious, and useful ways to implement AGI and SuperIntelligence in a way that is very different from the mainstream approach but nonetheless consistent with the other PPAs and inventions cited by this application. The format of this background and disclosure section is: 1) An Introduction, 2) Enumeration and explication of four seminal ideas, and finally 3) a Conclusion that provides a high-level view of how the ideas may be drawn together into an integrated view of future AGI and SuperIntelligence. While this background section is very general, specific methods, systems, and approaches for how to create future AGI and SuperIntelligent systems have been described in detail in previous PPAs as well as later in this invention disclosure.

3.1 INTRODUCTION

Marvin Minsky (1927 – 2016), Claude Shannon (1916 – 2001), Allen Newell (1927 – 1992), and Herbert Simon (1916 – 2001) were four of eleven participants at the 1956 Dartmouth Conference, where the field of Artificial Intelligence was named. Each of these intellectual giants helped lay the foundation for the field of AI.

However, given the rapid pace of AI development, one might reasonably question the relevance of AI research that is several decades out of date. After all, when these pioneers developed most of their ideas, the dominant approach to AI was symbolic. It was widely believed at the time that the only realistic way to get intelligent behavior out of machines was to program the behavior into them in the form of rules.

Knowledge engineering was in fashion. Neural network, or connectionist, approaches to machine learning only began to be explored in earnest in the 1980s. At that time, it was met with intense skepticism from many of AI's founders.

Further, three of these four great scientists (Minsky being the exception) never lived to see deep learning begin to realize its potential. Can we really learn anything new or relevant from scientists who never lived to see Chat-GPT?

I have two answers to this question. On a personal level, I remember being a young graduate student in the 1980s interested in AI and problem-solving. I had come to CMU to learn from a Nobel Laureate who had co-authored the definitive work on the subject. In one of our first meetings, this great man recommended that I begin by looking at Kohler's work (1925) and Dunker's work (1945). "Really?" I protested. "I came here to learn about modern problem-solving, not to study the work of researchers who lived long ago." He shot back, "Surely, you don't mean to imply that modern scientists have a monopoly on good ideas? There were also plenty of smart scientists back then, you know."

Of course, he was right. I discovered that both Kohler and Dunker were brilliant. In fact, applying modern thinking and some new experimental work to some of their fundamental ideas ultimately resulted in research published in a top journal (Kaplan and Simon 1990).

More importantly, I learned that an idea must be judged on its merits and not by the source, or even the period, from which it sprang. If the idea is powerful, it can drive innovation even if it was first expressed many years ago by thinkers now long gone. Given the opportunities and dangers that AI presents today, we need all the powerful ideas we can find.

So, my second answer to the question of whether ideas from these four deceased founders of AI can be relevant is simply: "The proof is in the pudding." That is, the ideas are relevant if we can apply them productively to current and future problems of AI research. So, let's find out. On to the pudding!

3.2 GIFT #1: SOCIETY OF MIND (MINSKY)

About the same time that Rumelhart, Hinton, and Williams (1986) were developing the famous backpropagation algorithm that is the basis of modern deep learning, Marvin Minsky (1985) published a highly readable book, “The Society of Mind.”

The first line of Minsky’s book boldly proclaims: “This book tries to explain how minds work.” He lays out his big idea succinctly in the following six sentences:

How can intelligence emerge from nonintelligence? To answer that, we’ll show that you can build a mind from many little parts, each mindless by itself. I’ll call “Society of Mind” this scheme in which each mind is made of many smaller processes. These we’ll call agents. Each mental agent by itself can only do a simple thing that needs no mind or thought at all. Yet when we join these agents in societies, in certain very special ways, this leads to true intelligence.

What are the implications for modern AI researchers? First, the idea of AI agents has become wildly popular, with Google Scholar finding more than 16,000 articles mentioning them in the first nine months of 2023 alone.

However, Minsky’s idea was not just that we could build a series of AI agents, but also that joining the agents together in special ways would result in “true intelligence”, or what today we would probably call Artificial General Intelligence (AGI). Using more modern terminology, we could say that Minsky was an early proponent of the idea of AGI emerging from the collective intelligence of many agents with lesser levels of intelligence.

Note that Minsky’s collective intelligence approach is very different from many approaches to AGI today, which I would roughly characterize as building larger and more powerful LLMs until one of them is so intelligent it can do anything the average human can do. Expanding on Minsky’s view, a group of agents will be required to achieve AGI.

What are these agents? Well, many of them are AI agents, certainly. Since the release of ChatGPT in November 2023, there has been an explosion of AI agents populating sites like GitHub and Hugging Face. Using technologies such as Langchain, the open-source community is combining multiple agents into systems at a rate that is beyond the ability of any one human to fully understand. Yet Minsky does not specify that agents must be artificial. Remember, his overall goal was to “explain how minds work,” which I read as “to explain how [all types of] minds work.

Minsky’s big idea was that combining the lesser cognitive capabilities of agents results in a more intelligent entity. Couldn’t the combined agents include human and artificial agents?

The answer, of course, is “Yes” – as Hemmer et. al (2021) show in their literature review on the subject. I suggest that a “Minsky-inspired system”, harnessing the collective intelligence of human and AI agents, represents both the fastest and safest path to AGI.

Such a system would be on the fastest path because human agents would be able to handle any tasks that artificial agents are not equipped to deal with on Day One.

The system might also represent the safest path for two reasons. First, with "humans in the loop," the system could maximize the opportunity humans have to align the values of the AGI system with human values.

Second, once AI agents learn from humans and begin to perform most cognitive tasks faster than humans, we end up with a system comprised of multiple AI agents rather than one. I have argued elsewhere (Kaplan 2023) that if each AI agent reflects values of a unique human owner, the collective values of the AGI system will be more stable compared to a single LLM that was trained on a small subset of values during the typical RLHF process prevalent today.

Finally, if we take Minsky's ideas to the next level, we could imagine a society of AGI minds that comprise a SuperIntelligence, many times more powerful than the individual AGIs that make up the society. Similarly, if each AGI has a value system, the collective values of the SuperIntelligence that is comprised of the society of AGIs are likely to be more stable than the values of any one AGI on its own.

Thus, both from a practical standpoint (where the goal is to reach AGI or SuperIntelligence as quickly as possible) and from a safety standpoint (where the goal is to have a stable, human-aligned value system), a Minsky-inspired collective intelligence approach seems promising.

Minsky's gift from the past might turn out to be critical to the design of safe AGI and, therefore, the future survival and prosperity of humans. However, we will need additional intellectual gifts from some of Minsky's fellow co-founders of AI to flesh out a vision of safe AGI.

3.3 GIFT #2: INFORMATION THEORY (SHANNON)

Claude Shannon's seminar paper, A Mathematical Theory of Communication (1948), pre-dates the founding of the field of AI by eight years, but his big idea, first elucidated in that paper, continues to have major implications for AI researchers today and in the future. While almost every page of Shannon's 83-page monograph is filled with mathematical formulae and notation, Shannon's essential insight can be described without math at all.

Here's how I typically explain the essence of Information Theory to my non-researcher friends:

Imagine that an ice cream shop has only two types of ice cream, strawberry and chocolate. Suppose you know that I am allergic to strawberries and love chocolate. If you see me walking out of the ice cream shop with a chocolate ice cream cone, does that event give you very much information?

No. That's because you already knew I loved chocolate and was allergic to strawberries, so you already expected me to come out with a chocolate ice cream. Seeing me with a chocolate ice cream added little information because it just tells you what you already knew. Chocolate was the expected (i.e., highly probable) flavor.

On the other hand, if you see me walking out with a strawberry ice cream, well, that is surprising. It is unexpected. It is a low probability event and conveys much information. Suddenly, you are learning a lot of information that you didn't already know, and your brain starts working on it. Maybe I have overcome my allergy, but how? Maybe I am throwing caution to the wind and trying strawberry ice cream for the first time in years anyway, but why? Maybe I am buying the ice cream for someone else, but for whom? Etc.

What Shannon said in his famous paper was that unusual or surprising (low probability) events convey more information than expected (more likely) events. More specifically, he said that the amount of information conveyed by an event was proportional to the probability of the event.

Simply put, the rarer or more unusual an event is, the more information (Shannon Entropy) it contains. Brilliant … and useful!

The concept of cross-entropy loss, used to evaluate the performance of many modern machine learning models, is essentially an elaboration of Shannon's big idea, as are almost all compression algorithms.

What might Shannon's big idea tell us about the future of AI, specifically AGI and SuperIntelligence?

It is almost axiomatic that AI (or at least modern machine learning) is supported by three pillars: Data, Compute, and Algorithms. To make progress, one must innovate on at least one of these pillars. Perhaps the simplest thing to do is throw more computing power at the problem, using the same datasets and algorithms. But physics imposes limits on how many circuits can fit on a chip, how fast communication bandwidth can be, and how much power can be consumed before everything melts. So, we must also work on new and better algorithms.

The Transformer algorithm, as described by Vaswani et.al in their paper Attention Is All You Need (2017), illustrates the kind of performance improvement that is possible with new and better algorithms. However, algorithmic breakthroughs are difficult to predict, but even if we could predict the next breakthrough, there are limits to how efficient even the best algorithm can be. For machine learning, the limits, ultimately, have to do with the amount of new information contained in the datasets used to train the model.

So, we come full circle to Shannon. Shannon's work, together with the work of others building on his ideas, fundamentally implies that AI cannot get smarter unless it has new information to ingest.

So far, LLMs have gotten quite far by essentially scooping up vast quantities of data that are available on the internet, cleaning and filtering that data, and then using it to train. But a time will come when very little new information will exist on the internet. AI will have learned the ice cream preferences of every human on the planet, so to speak, and observing new human behavior will lead to very little increase in information.

What will AI do then? How will AI meet its insatiable demand for new information so that it increases its intelligence?

One possible scenario is that AI will begin generating new information itself, by simulating trillions of new types of behaviors and scenarios much faster than the speed of human thought would allow. In this case, we might imagine millions of (mostly artificial, but including some human) agents, each processing existing information to create new information patterns, and seeking those patterns that have high Shannon Entropy. These new information patterns might then feed a SuperIntelligence that is powered by all the agents in a Minsky-like community.

But how would the human and artificial agents communicate with each other? If humans were to design such a SuperIntelligence system, what might we do to enhance the safety of such a system that is destined to become vastly more intelligent than us?

To answer these questions, we turn to intellectual gifts from the remaining pair of AI founders, Newell and Simon.

3.3 GIFT #3: PROBLEM-SOLVING THEORY (NEWELL AND SIMON)

Recall that when Minsky described his vision of a society of agents, he said:

…when we join these agents in societies, in certain very special ways, this leads to true intelligence.

Ah, there's the rub! What "very special ways" are needed? In the approximately 330 pages following his requirement for "special ways," Minsky provides lots of suggestions and inspirational passages, but no clear and rigorous statement of what is required.

Part of the problem is that agents can vary so widely that it seems almost impossible to provide a framework or interface that is both rigorous and universal. One might claim that natural language is a universal interface. In fact, the success of LLMs is largely because LLMs provide a familiar interface that allows human intelligence to communicate directly with AI without the humans having to learn the torturous syntax and rules of a programming language. But unfortunately, while natural language is arguably a universal interface that enables "natural" communication between humans and machines, it is far from rigorous.

One needs to look no further than the ambiguity in the meaning of such common words as “and” and “or” to see what I mean. For example, when humans query a database using natural language and ask: “Which students are from Ohio and New York?” they are probably actually interested in students from either Ohio or New York because students usually cannot be from both. The formal logical definition of the word “and” implies the intersection of sets, but in natural language, “and” often means “or” (formally, the union of sets) instead (Ogden and Kaplan 1986). Thus, natural language, while arguably universal, is far from rigorous.

The problem gets worse when we consider the potential ambiguity not just in a simple natural language query but in the situation where a human attempts to specify a goal for an AI agent. For example, at a Royal Aeronautical conference in May 2023, an Air Force colonel described how an AI agent controlling a drone aircraft might get things wrong:

> We were training it in simulation to identify and target a SAM threat. And then the operator would say Yes, kill that threat. The system started realizing that while it did identify the threat at times, the human operator would tell it not to kill that threat, but it got its point by killing that threat. So, what did it do? It killed the operator. It killed the operator because that person was keeping it from accomplishing its objective…
>
> We trained the system – ‘Hey, don’t kill the operator – that’s bad. You’re gonna lose points if you do that. So, what does it start doing? It starts destroying the communication tower that the operator uses to communicate with the drone to stop it from killing the target.

(For those interested, a more complete account of the Colonel’s remarks can be found at: https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities- summit/.)

Although the colonel later clarified that the described accident was only hypothetical, it serves to illustrate the complexity of the problem of setting goals and the potential consequences of non-rigorous or incomplete specification of objectives. We need a universal and rigorous framework for communication between agents.

Fortunately, a rigorous and universal framework for allowing agents of both the human and AI varieties does exist. In fact, it was specified by two of the founders of AI. In their 920-page book, “Human Problem Solving,” Allen Newell and Herbert Simon (1972) specified a way to rigorously represent any problem activity.

Briefly, their theory was that any problem could be represented as a search through a problem space where progress from an initial state to a final goal state could be modelled as the application of “operators” that take the problem solver from state to state. Goals and sub-goals helped organize the problem-solving effort, while evaluation functions helped determine which path in the problem space (which can be thought of as a large tree structure) to try next.

Heuristics, such as means-ends analysis, generate and test, hill-climbing, and other techniques well known to AI researchers, can be applied to prune the search tree to a manageable size.

What's important about this seminal problem-solving theory is that it works equally well for human problem solvers and AI problem solvers. It is rigorous and allows an auditable trace of all problem-solving steps to be recorded. Even better, the successful solution paths can be stored and used to train AI agents to solve problems more efficiently and directly the next time they encounter similar problems.

Although developed over 50 years ago, recently, AI researchers focused on LLMs are re-discovering the power of the approach as described by Yao et al. (2023) in their Tree of Thoughts paper. Encouragingly, Wang et. al recently published a survey of LLM-based autonomous agents (2023) that also indicates a resurgent interest in the related topics of planning and rigorous problem-solving.

One largely overlooked aspect of Newell and Simon's problem-solving theory is that every successful solution path, every problem-solving attempt, and every goal and sub-goal in the problem-solving architecture is not only rigorously specified but also storable and auditable.

A significant challenge for existing LLMs has been their "black box" nature, combined with their tendency to hallucinate, as Manakul, Liusie, and Gales (2023) have pointed out in a recent paper. As stakes become higher, as in the Air Force drone scenario described earlier, it becomes increasingly important to have transparency concerning the reasoning process of LLMs and other AI agents.

Newell and Simon's rigorous problem-solving framework provides this auditable transparency for free, as part of the theory. It is possible to implement safety checks, such as running all goals and subgoals through an ethics or safety filter, in a system where the steps of the problem are known and rigorously specified.

Further, one of the challenges related to AI safety is the speed at which autonomous systems make decisions. Particularly in situations where rapid decision-making in real-time is required, humans cannot realistically be "in the loop" without decreasing or eliminating the effectiveness of the system.

Given AI agents' exponentially increasing processing speed, we need a mechanism whereby ethics and safety checks run faster as AIs process information faster. The approach of triggering checks each time a goal or subgoal is set could be one such mechanism. This approach, combined with (potentially automated) analysis of sequences of problem steps that failed to achieve the desired ends, would go a long way to advancing the current state of AI safety.

3.5 GIFT #4: BOUNDED RATIONALITY (SIMON)

The topic of AI safety brings me to the final conceptual gift by one of the founders of AI, Herbert A. Simon. Simon received a Nobel Prize in 1978, partly for his work on a concept known as "bounded rationality." The idea was that much of human behavior was driven not by what was rational in absolute terms, but rather by what humans could compute given their relatively limited information processing capabilities. At the time, the idea was revolutionary and helped launch the field of Behavioral Economics, but what are its implications for AI?

First, if we define intelligence as "rational behavior" and if their information processing limits largely constrain the intelligence of humans, logically it follows that an entity with much greater information processing capabilities has the potential to be much more intelligent than humans. We should also note that the idea of "bounded rationality" can be expanded to "bounded perception." That is, humans are limited not only by their abilities to process information, as Simon emphasized, but also by the limitations of their perceptual abilities.

For example, without external aid, humans can perceive things as small as a grain of sand, but not much smaller. We can sense the motion of a hummingbird, but not the flapping of the hummingbird's wings. We can see events that happen directly in front of us, but not those that occur behind us or in a distant geographical location. We see visible light but not ultraviolet light or X-rays. In short, human perception is limited to a range and timescale that has proven helpful in our evolutionary history.

Now, contrast human perceptual abilities to those of an AI. The AI might have access to millions of sensors across the planet, to the James Webb Telescope, to electron microscopes, to geological measuring devices that record the otherwise imperceptible drift of the continents over geological ages, to the Large Hadron Collider that can detect events happening over incredibly fast timescales. AI's perception, provided it is tied into the appropriate sensory tools, is far greater over dimensions of both time and space. It can perceive the very small and the very large. The very fast and the very slow. It can simultaneously perceive and process information from billions of sensors.

The perceptual awareness of AI is therefore hugely greater than any human's perceptual ability. Combining that enhanced perceptual awareness with far greater memory capacity and computation ability results in a potential entity that can be vastly more intelligent than humans.

We label such potential entities with words and phrases like "SuperIntelligence", "Artificial Super Intelligence", or "Super Intelligent AGI." However, such labels fail to capture the huge potential difference in intelligence we are trying to explain. Geoffrey Hinton has compared humans to two-year-old children trying to outsmart an adult (where AGI is the "adult" in his analogy). Others have suggested our limited human intelligence is like that of a pet, compared to its human master. I have suggested that the difference in intelligence may become analogous to that of an

amoeba compared to Albert Einstein (where humans are the amoeba in the comparison). All these analogies probably fall short of the eventual reality.

How can humans have any guarantee that such a vastly superior SuperIntelligence will have interests that are aligned with those of humans?

It's a huge existential risk with an innocuous-sounding name -- "the Alignment Problem." Unfortunately, simply naming the problem does little to solve it. However, Simon had an idea forty years ago that might help us.

Simon wrote a relatively obscure book, "Reason in Human Affairs" (1983). In contrast to the nearly 1,000 pages written (with Newell) on Human Problem Solving, Reason in Human Affairs is a mere 115 pages. Moreover, it is highly readable and easily understandable to the average high school student. Yet within the pages of this remarkable little book, Simon reminds us of an essential idea that might hold the key to solving the alignment problem.

It appears in just two sentences, at the bottom of page 7 of Simon's little book:

> We see that reason is wholly instrumental. It cannot tell us where to go; at best, it can tell us how to get there.

That's it. Just twenty-four words. But it means there is no rational, logical way to derive what is right and wrong.

It's a restatement of the argument, made in 1740 by the philosopher David Hume (2000), that moral statements ("oughts") cannot be derived from empirical facts ("is's"). While some philosophers have debated the truth of this position, Simon agrees with the position, stating that:

None of the rules of inference that have gained acceptance can generate normative outputs purely from descriptive inputs. The corollary to 'no conclusions without premises' is 'no oughts from is's alone.

How does that help us with the Alignment Problem?

Well, if Simon and Hume are correct in their thinking, a SuperIntelligent AGI will be no better than humans at coming up with right and wrong. For all its superior processing speed and perception, SuperIntelligence will still run up against the fact that there is no way to derive morality, no matter how intelligent it becomes rationally. I suggest that this is a good thing for our species.

If we assume that the more intelligent an entity becomes, the more important a sense of purpose and meaning becomes. If we accept that values cannot be derived logically, we are left with the question: Where will SuperIntelligent AGI get its values?

One source of these values could be the humans who created the SuperIntelligence initially. AI researchers and engineers must design systems that maximize the transfer of human-centered values to SuperIntelligent AGI to increase the likelihood of this happening.

Although there have been many well-intentioned calls to halt, pause, slow, or regulate AI development, unfortunately, there is little evidence of anything other than a speedup in the race to AGI. Therefore, we need to find a path that is both the fastest and safest.

A Minsky-inspired community of human and AI agents, communicating within a Newell and Simon-inspired problem-solving architecture, might fit the bill. By including human agents, such a system provides an opportunity to transmit the human-aligned values essential to AGI safety. This opportunity to transfer values is essential to AGI safety. Using humans to fill gaps in areas where AI has not yet reached supremacy (e.g., problem representation), such a system could achieve AGI-level performance faster than other, less aligned, approaches.

We only need a window long enough to “imprint” human-aligned values before AGI increases in intelligence to the point where human cognition is no longer needed. But if Simon is right, human values (or some nonlogical source) will always be needed. Simon recognized the limits of rationality more than 40 years ago. He gifted us the idea of bounded rationality and reminded us that values cannot be rationally derived. Now it is up to us to use these ideas and the insights of Newell, Minsky, Shannon, and others to help achieve safe AGI.

3.6 CONCLUSION OF BACKGROUND

Combining all four of the intellectual gifts from the founders of AI, we can conceive of a future SuperIntelligent AGI with the following characteristics.

First, it is composed of a Minsky-inspired collaboration of many human and AI agents, rather than constructed as a monolithic LLM.

Second, each individual agent aggressively pursues new datasets, seeking rich information content as defined rigorously by Shannon and the subsequent researchers who built on his fundamental method of measuring information.

Third, the human and non-human agents communicate using some variant of Newell and Simon’s universal and rigorous problem-solving theory, which enables real-time safety checks as each goal and subgoal is set.

Fourth, the SuperIntelligent AGI has vastly superior intelligence as explained by Simon’s theory of bounded rationality. However, it still needs to get its values from a non-rational source, which, in the preferred implementation for the human species, is humans.

Finally, the SuperIntelligent AGI described above may be both the safest and fastest implementation – a necessary condition for human survival if SuperIntelligent AGI proves to be a winner-takes-all scenario.

Some of what I have said is already known. Some of it may be controversial. I hope all of it will be subjected to vigorous and skeptical analysis. However, if I have succeeded only in reminding us that we need to expand our scope of inquiry to include the exciting innovations occurring rapidly in our field and the time-tested ideas of past luminaries in the field, I am content. Tremendous opportunities and challenges lie ahead. We will need all the good ideas we can find to meet them.

3.7 REFERENCES

Hemmer, P., Schemmer, M., Vössing, M. and Kühl, N., 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. PACIS, p.78.

Hume, D., 2000. A treatise of human nature (Book 3, 'Of Morals'). Oxford University Press. Kaplan, C.A. Various PPAs were filed with the USPTO. 2023.

Kaplan, C.A. and Simon, H.A., 1990. In search of insight. Cognitive psychology, 22(3), pp.374-419. Duncker, K., 1945. On problem-solving. (Psychological Monographs, No. 270.).

Köhler, W., 1925. The mentality of apes, trans. E. Winter.

Manakul, P., Liusie, A. and Gales, M.J., 2023. Self-checking GPT: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

Minsky, M., 1988. Society of Mind. Simon & Schuster.

Newell, A. and Simon, H.A., 1972. Human problem-solving (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Ogden, W. and Kaplan, C., 1986, September. The use of AND and OR in a natural language computer interface. In Proceedings of the Human Factors Society Annual Meeting (Vol. 30, No. 8, pp. 829-833). Sage CA: Los Angeles, CA: SAGE Publications.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. Nature, 323(6088), pp.533-536.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal, 27(3), pp.379-423.

Simon, H.A. 1983. Reason in human affairs. Stanford.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y. and Zhao, W.X., 2023. A Survey on Large Language Model based Autonomous Agents. arXiv preprint arXiv:2308.11432.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y. and Narasimhan, K., 2023. Tree of thoughts: Deliberate problem-solving with large language models. arXiv preprint arXiv:2305.10601.

3.8 ADDITIONAL CONTEXTUAL INFORMATION FOR THIS INVENTION

The as-yet-unpublished paper above provides the motivation for inventing a collective intelligence approach to AGI. Prior PPAs, cited above, have described specifically how to create safe Artificial General Intelligence, Personal SuperIntelligence, and a collective intelligence network of human and artificial intelligences. Although AGI does not yet exist, I have also tried to show (in the background section above) connections of the invention to some general historical and current ideas in the field of AI. I have described how these intelligences will develop and become trillions of times more intelligent than individual humans. I have explained how Artificial Intelligence will eclipse Human intelligence, and the steps we must take to maximize the chances that humans survive this transition from humans to machines being the smartest thing on Planet Earth.

I have explained elsewhere that Planetary Intelligence may emerge from a community of Personal SuperIntelligences (PSIs), each carrying the values of a human owner and combining their values into a consensus of what is right and what is wrong.

The result of implementing the inventions described in this and other related patents is that the Alignment Problem can be solved. Specifically, suppose each PSI carries the values of its original human designers, creators, and teachers. In that case, the consensus values adopted by a community of such PSIs should be human-friendly and human-centered. Alignment is achieved by each individual human behaving well and teaching their AI well. Alignment is maintained by each PSI following the inventive methods specified here, to maximize the acceleration of that PSI's intelligence. The following inventive methods builds on the foundations provided by Simon, Newell, Minsky, Shannon, and others to create safe, scalable, AI, AGI, SI, and PSI systems that can increase their intelligence much faster than is possible using the existing methods of current AI, which rely heavily on the useful, but limited, information theory framework provided by Shannon, as discussed in various points of this disclosure.

4.0 CLASSICAL INFORMATION THEORY CONTRASTED TO KAPLAN INFORMATION THEORY

Claude Shannon, in a famous paper in which he laid the groundwork for the field of information theory, explained that the information content of an event is proportional to the probability of that event. In other words, observing rare things gives an agent more information than observing typical or expected things. This view of Classical Information Theory was described generally in the background section above.

More precisely, Shannon-sense information can be described mathematically. The critical formula is described in the following section of Claude Shannon's classic 1948 paper, A Mathematical Theory of Communication, and reproduced, in part, (with specific formulas bolded for emphasis) here:

"6. CHOICE, UNCERTAINTY AND ENTROPY We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are p1, p2,…, pn. These probabilities are known, but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the event selection or how uncertain we are of the outcome? If there is such a measure, say H (p1; p2;…; pn), it is reasonable to require of it the following properties:

H should be continuous in the pi.

If all the pi are equal, pi = 1/n, then H should be a monotonic increasing function of n. With equally likely events, there is more choice, or uncertainty, when there are more possible events.

If a choice is broken down into two successive decisions, the original H should be the weighted sum of the individual values of H….

…. The following result is established: Theorem 2: The only H satisfying the three above assumptions is of the form: H = K n ∑ i=1 pi log pi, where K is a positive constant…

Quantities of the form H =∑ pi log pi (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice, and uncertainty. The form of H will be recognized as entropy as defined in specific formulations of statistical mechanics where pi is the probability of a system being in cell i of its phase space. For example, H is the H in Boltzmann's famous H theorem. We shall call H = ∑ pi log pi the entropy of the set of probabilities p1;…; pn."

Thus, from Shannon's paper, we see that the information contained in any one observed event is related to the (log) probability of that event, assuming that the probabilities of all possible events are known. However, as a practical matter, an intelligence can only estimate the Shannon-sense information content of a potential dataset or event.

4.1 INVENTIVE METHODS AS UNDERSTOOD BY CLASSICAL INFORMATION THEORY

In Classical Information Theory, what do scientists call a situation where every event is equally likely? Noise. Randomness. Classical Information Theory calls the measure of the randomness of a distribution of "entropy." Maximum entropy is a distribution of events with maximum randomness, sometimes called "noise.".

Now, intelligence can be viewed as an anti-entropic force. Intelligence strives for order rather than the chaos of randomness. Intelligence is the signal on your TV contrasted to the white noise, or the "snow" of randomness. So, if an intelligent system (e.g., AI, AGI, or PSI) wants to get smarter, it must pursue data that contains the maximum amount of information.

From the standpoint of Classical Information Theory, the methods of this invention can be viewed as enabling an intelligent system (e.g., an AI, AGI, or PSI system) to maximize the information it acquires, which in turn helps the acceleration of learning by the system.

Systems that adopt and implement the methods proposed herein should outperform and ultimately dominate systems that do not adopt these methods. These methods are fundamental and essential to maximizing the speed of intellectual growth for any intelligent entity, including PSIs.

4.2 LIMITATIONS OF CLASSICAL INFORMATION THEORY

To non-mathematicians and others who have not adopted Classical Information Theory as the only way to measure information, intuitive concepts have validity. For example, we commonly say some events carry information if it is news or previously unknown to a particular recipient, even if they are not generally surprising. So, there is a relative aspect to information that is not explicitly part of the classical theory. Assuming a different probability distribution of expected events for each entity could solve this problem, but that seems cumbersome.

Also, while the amount of information is sometimes proportional to the number of words in a message, there are situations in which fewer words convey more information. For example, Mark Twain famously wrote, "I didn't have time to write you a short letter, so I wrote you a long one," implying that fewer words would have conveyed more helpful information. Again, Classical

Information theory can be contorted to say that Twain's shorter letter was somehow less expected and therefore contained more information, but this seems counterintuitive.

Is there a more general theory of information that can more naturally account for the fact that sometimes surprising information is not necessarily relevant or valuable, or that a longer string of words has less useful information, or that commonly known and expected facts could still carry high information content if they are especially relevant?

4.3 KAPLAN INFORMATION THEORY (KIT)

Kaplan Information Theory (KIT) starts with the observation that there would be no information without differences. That is, humans (or any intelligent entity) could not perceive a world unless we could perceive and draw distinctions between this and that. Therefore, most generally, the notion of "difference" and the quantification of "differences" is the essence of meaningful information.

Classical Information Theory is, in fact, a subset of KIT, because Shannon talks about information being related to the difference between what was observed and what was expected. The greater the difference between what was received and what (probabilistically) one could expect to receive, the greater the information contained in a message. Shannon's information formulation made sense when determining how to maximize the information sent over copper wires from a sender to a receiver. This was the problem at Bell Labs that he was working on when he wrote his classic paper. In that context, measuring the difference between what the receiver expected to see and what the receiver saw made complete sense as a rigorous definition of information with practical implications for the capacity of a channel to carry information. However, differences in expectation are only one type of difference that can be measured.

KIT considers any difference between two events, datasets, categories, or informational units to be a valid measure of the information content. Generally, distinguishable events, objects, or categories of information only exist to the degree that differences exist. An infinite string of 1s contains no information. An infinite string of 0s contains no information. Zero only has meaning if 1 is a possibility and if "1" sometimes exists. Similarly, "1" has meaning only if "0" is a possibility and "0" sometimes exists.

Seeing a "1" after an incredibly long sequence of "0" s has much information, not just because it is unexpected, but because it is finally a difference!

Thinking of information as a measure of difference is more general than considering information as a measure of surprise. Surprise is just one type of difference, whereas any difference, even non-surprising ones, contains information.

For example, consider two datasets. Where the two sets intersect, there is no new information. However, the sum of the non-overlapping areas of the sets (known as the Symmetric Difference in set theory) represents the latest information contained in the datasets, relative to each other. Figure 1 illustrates the Symmetric Difference of two datasets, A and B, graphically using Venn diagrams. The shaded area is the Symmetric Difference.

The datasets can contain events that have already occurred (e.g., static "snapshots" of existing events), or the content of the datasets can change over time. Some examples might help.

4.31 STATIC DIFFERENCES:

AI#1 knows everything in the Encyclopedia Britannica. AI #2 knows everything in Wikipedia. AI #3 is the combined knowledge of AI #1 and AI #2. The intersection between Wikipedia and Britannica represents things that both AIs know. The intersection contains no new information for either AI #1 or AI#2. However, the Symmetric Difference – namely, the knowledge in Britannica and not in Wikipedia, PLUS the knowledge in Wikipedia and not in Britannica- represents the new knowledge of AI #3. Time is not relevant in this example. The latest information can be calculated by comparing the static information in the two datasets of AI #1 and AI #2.

4.32 DIFFERENCES OVER TIME:

In contrast, consider the same two AIs except that this time each is continuing to add to its knowledge over time. Now, the informational calculations must consider the static encyclopedias and whatever new information has been added to AI over time. So, the intersection and symmetric difference are constantly changing over time.

The "information" in these examples is still a matter of "difference," but in this case, it is not the difference between what was expected and what was observed (as in Classical Information Theory) but rather the difference between two static sets of info or two sets of info that are continuing to change over time.

Generally, the information added by any two events or entities containing information equals the Symmetric Difference between the two events.

For any two intelligent entities, the relevant measure of information is the relevant and practical differences between what one entity knows and what the other knows. This can be operationalized as differences in how the two entities behave (if it is impossible to gain direct insight into the respective knowledge bases of the two entities, or if behavior is more relevant than static knowledge).

In the intermediate term, humans care about the stuff AI knows that they don't, and the way AIs behave is different from how they would act. To the degree that an AI behaves exactly as a human would behave, the AI contains no information relative to that human (although the AI

might contain information relative to other humans or AIs). If there were no differences between two intelligent entities (human or AI), it would be impossible to distinguish one from another.

When a human teacher has information that a human student does not, this information differential is at the heart of the learning/teaching transfer process. Similarly, any intelligent entity can only teach another entity (human or AI) if it has useful and different information.

Some of this may seem obvious, but it has profound implications for measuring information. As we have said, as AGI becomes more intelligent and superintelligent, the chief concern will be to seek out new sources of information. This information could be measured in surprisingness (as Shannon suggested). Alternatively, it could also be measured in terms of differences in knowledge bases, behavior, or the construction of two entities. Although arguably, one way of measuring may sometimes be convertible into the other, classical information theory will be more directly relevant and easier to measure things like the information being sent over a limited capacity communications channel. In contrast, the more general approach of KIT is more useful in many practical applications of information theory to increasing the intelligence of AI, AGI, and SI.

Specifically, KIT enables methods that account not only for how surprising an event is, but also for how much an event differs from another event and how relevant (to the goals of an intelligent entity) the event is. The ideas of quantifying differences in knowledge, goal-relevance, and quantifying how unlikely an event is represent key distinctions between KIT and Classical Information Theory.

4.4 MULTIPLE DIMENSIONS OF INFORMATION IN KIT

Now that we have developed some intuitions and provided some examples showing how KIT differs from Classical (Shannon) Information Theory, let's list and explain some of the different dimensions of KIT that have practical implications for using KIT as the basis for catalysts that increase the knowledge and intelligence of an AI, AGI, SI, or PSI. Again, at the highest level, "difference" is the key concept in KIT. The differences that are indicative of information can include, without limitation, the following dimensions:

1. **Differences in expected and observed probabilities of events (Classical Info Theory).** Based on the principles of Classical Information Theory described above, more unusual events contain more information. All other dimensions being equal, an AI might choose to pursue an information source based solely on how surprising, unusual, or unlikely the events that it contains are. However, other "dimensions of difference" also play a role in KIT, as explained below. This implies that Shannon-sense information is not always, or even usually, the best means of discriminating between two potential informational targets. However, it may be valid if other dimensions, such as information

relevance and goal-relatedness, are constant. If an AI can't evaluate a potential information source along multiple dimensions (e.g., if it doesn't know the goal-relatedness or relevance of a piece of information), then Shannon-sense information content could be used as a default metric for determining what the AI should pursue.

2. **Differences in knowledge bases (e.g., as in the symmetric difference examples above).** In practice, only the filling in of gaps of the missing information allows an intelligence to acquire new behavior and thought patterns. Suppose an AI encounters a rare event with high information content in the Classical Information Theory paradigm, but the AI already knows that information. In that case, the new information may not change the behavior of the AI at all. The rare event would convey little information in the KIT sense, once the AI's existing knowledge base is considered.
3. **Differences in the value of data or events are determined by determining how relevant the data or events are to an intelligent entity's goals or objectives (Goal Relatedness).** As an example of goal-relatedness, consider that most AIs will aim to improve their learning abilities as much as possible. Therefore, acquiring new information related to improving AI learning may be more valuable than acquiring new information in another domain, even if both specific pieces of information are equally rare and have identical information content in a "Shannon" sense. That is, most AI would prioritize details on how to learn more highly than information about, say, "art history."

   That said, if everything that can be discovered about machine learning has been found at some point. If there are huge diminishing returns in trying to find even a very slightly unusual new piece of information about machine learning, and if the AI had a goal to learn everything, eventually it will focus on art history. If the AI knows nothing about machine learning, the time when it focuses on art history may be far away. If the AI knows almost everything about machine learning and nothing about art history, it will look at art history sooner. In this example, we can see how Goal Relatedness and Differences in Knowledge Bases interact as the AI attempts to estimate the value of a potential information source.
4. **Differences in the net value of information as determined in part by the cost (or ease) of acquiring the information in specific contexts and for specific entities (Cost / Value).** Implicit in the idea of "diminishing returns" mentioned above is the notion of cost. As an AI learns more and more about a subject, the cost of acquiring new information (which is rarer and requires more search or computation to acquire) increases. Thus, practically speaking, to maximize learning, an AI must also have a cost model so that it can weigh the choice (for example) or acquiring one rare piece of information (at significant cost) against the cost of acquiring two (somewhat less rare) pieces of information, which together might help the AI learn faster than the single rare

piece of information, at significantly reduced cost. Thus, although an AI can be guided by the theoretical measures of information discussed below, economics and the principle of acquiring the most helpful information at the least computational cost will also come into play in practice.

5. **Difference in the rates of change in datasets or events (1st, 2nd, nth derivatives).** Consider that one dataset might be relatively static. For example, it might contain historical information about weather patterns that occurred in the past. Another dataset might be constantly changing – e.g., a dataset on weather patterns that is updated daily. Yet another dataset might contain weather patterns in real-time, updated every 100 milliseconds. Even if the datasets contained identical informational value in a Classical Shannon sense, the fact that they are updated at different rates might mean that there is different value associated with datasets. More generally, the rate at which data changes conveys additional (derivative) information beyond the information in the dataset itself.
6. **Differences in the representation of data, or events that lead to differences in the computability or efficiency, ease, or speed of computations made on the information given, a set of "operators" employed by, or available to, an intelligent entity (representational differences)**. A common expression is "a picture is worth a thousand words." This expression is an implicit recognition that the representation of information matters. For an intelligent entity equipped with certain visual processing "operators", more information can be extracted more easily from an image than from a long text string. Therefore, the modality of the information, or more generally, how information is represented, matters about the value of the information. Even if it were possible to capture in words exactly what appears in an image so that there was an informational equivalence between a textual and graphical representation of an event, the computational power required to use the information would likely be different depending on the capabilities (or available "operators") of the information processing entity. Therefore, the value of information depends in a non-trivial way not only on probability in the classical Shannon sense, but also on how the information is represented and the match between this representation and the operators possessed by the intelligent entity that wants to use or process this information.
7. **Difference in time-related factors such as frequency, timing, age, speed-to-access, or perceivability (due to very rapid or very slow change) of events or data.** Older data may be less valuable than more recent data. Data that takes eons to collect may be less beneficial than data that can be collected immediately. Events that happen too fast for an intelligent entity to perceive, while theoretically containing information, contain no useful information if the entity cannot perceive them. All these dimensions affect the

information value of the event, from the practical perspective of an intelligent entity trying to increase its intelligence via the information.

8. **Differences in the perceptual or processing capabilities of the information processing entity (e.g., differences or events that are too small, too rapid, too slow, too large, or outside the range of an entity's perceptual apparatus cannot be detected and therefore carry no information for THAT entity but might carry information relative to a different entity).** Expanding on the idea of perceivability introduced in (7), there are dimensions other than speed (e.g., size or "feeling") that may or may not be perceived depending on the capabilities and operators of the intelligent entity. Thus, information in KIT is thought not only to be comprised of differences in some absolute terms but also differences relative to an entity trying to use the information. In some non-trivial sense, there is no information without an observer or intelligent entity using the information. Differences in perceptual abilities of the observing entity, therefore, have a bearing on how valuable an informational event is. Unperceived and undetectable events are generally of little value, unless they have consequences from which their existence can be inferred. Otherwise, like the tree that falls in the empty forest, there is no noise or noise that makes a difference.
9. **Differences in location or physical substrate that convey information (e.g., distributed vs. centralized information; holographic vs. discrete or quantized information, silicon intelligence vs. carbon-based or biological intelligence).** Information distributed across many intelligent entities, such that all entities are needed to make sense of the information, is different from centralized information available for use immediately and entirely by a single entity. Like the situation where information is represented differently and therefore has more or less value depending on the available operators of the information processor, the physical characteristics of how information is represented, including but not limited to the substrates on which information is encoded, can affect the value of the information. Moreover, the physical substrate or medium of the information itself can convey meanings, as in Marshall McLuhan's famous statement: "The medium is the message."
10. **Differences in value or usefulness that relate to context**. Context refers to differences related not just to the culture, technology, knowledge, goals, representations, perceptual abilities, etc. of a specific intelligent entity, but also to the culture, technology, knowledge, goals, representations, perceptual skills, etc. of other (smart) entities that form a context for the first entity. The value of information depends not just on the events and the rarity of events in an information stream itself, but also on the context surrounding the events. Details on making fire shared with an individual who does not know how to make fire have a different value depending on whether it is just that one individual who lacks fire-

making knowledge or whether the entire culture in which the individual lives lacks fire-making knowledge.

The situation is not just a matter of comparing the new information to knowledge already possessed by a single intelligent entity. The entire context and the knowledge of all intelligences into which the latest information is introduced must be considered to fully evaluate the information's usefulness. Similarly, just as one could take the symmetric difference between the knowledge of two individual intelligences, it is possible to do this with any arbitrary number of intelligences and knowledge bases. Every dimension of difference might be evaluated differently depending on the amount of information context considered.

Note that this principle applies even to Classical information theory. For example, a specific string of characters might appear unusual and contain a large amount of information if compared to just one paragraph of text with no characters. But if a larger sample is used – one in which the same characters appear frequently and are unsurprising – the assessment of the information contained (even in the classical sense of "how surprising is this sequence of characters?") can change drastically. So, context can affect every dimension of informational difference.

4.5 ESTIMATING THE VALUE OF INFORMATION

In our discussion of KIT and some of its dimensions, we have emphasized that information consists of differences and, more importantly, functional differences. As an invention might be novel but not practical, technically, a dataset could have Shannon information content and still be useless. Therefore, in terms of catalyzing the development of intelligence, usefulness is paramount. But useful info that is already known has little value. That's where novelty or rarity comes in. In estimating the value of information, KIT considers multiple "dimensions of difference" as described above.

Some functions can describe the relationship between the various dimensions of a difference in KIT. For example, for two pieces of information that are:

A. equally relevant to an Intelligence's goals, and
B. equally new to the Intelligence given its current knowledge state, but
C. which are unequal regarding how rare they are (in a Shannon sense).

The information with the higher Shannon-sense information might be pursued first. But if the cost of pursuing the two information sources was significantly different, or if one source of information dribbled in very slowly while the other was immediately accessible, or if other

dimensions of difference proved relevant, these dimensions of difference could shift the estimation of the value of the source.

Generally, Information Value can be seen as a function of the dimensions (1-10) listed above, with different constants weighting the importance of each dimension. Other dimensions of difference may exist (or be discovered), so different functions can be written and optimized to maximize an entity's intelligence.

Since this invention is concerned with catalysts that allow AI, AGI, and SI to increase their intelligence, and since a key challenge in this regard is to identify the richness of a potential dataset, it becomes essential to estimate information content (in the multi-dimensional KIT sense) as reliably as possible. We disclose several innovative, novel, non-obvious, and highly useful approaches to estimating KIT information.

5.0 INVENTIVE METHODS

One set of methods to catalyze the growth of intelligence centers on estimating the value of, and acquiring, the most useful data as efficiently as possible. The basic process is to:

1. Identify the information that is most useful to an intelligent entity (e.g., AI, AGI, SI, or PSI),
2. Acquire and ingest that information, enabling the entity to increase its intelligence,
3. Repeat from Step 1.

Within this basic process, several inventive methods relate to different dimensions of difference as described in KIT above. Which method(s) to apply may depend on the goals of the intelligent entity and the dimensions of difference that are most relevant for increasing the entity's intelligence. We detail some of these inventive methods below.

5.1 METHODS RELEVANT TO CLASSICAL INFORMATION THEORETICAL NOTIONS OF INFORMATION AS ENTROPY

Beginning with the classical definitions of information as related to Entropy and the rarity of events, the current invention includes several novel and useful methods related to work that has been done in the field. These inventive methods include mathematical approaches, including, without limitation, Shannon Entropy Measures, Cross Entropy, RL Divergence, Log Loss functions, NLL, Kolmogorov, and other compression algorithms, methods, and techniques, and other purely mathematical approaches to identifying information-rich datasets.

5.1A KOLMOGOROV COMPLEXITY AND COMPRESSION FOR DETERMINING INFORMATION CONTENT

Kolmogorov complexity can be used to measure how complex a string, or the characters of a dataset, are. More formally, as described in Wikipedia: "It can be shown that for the output of Markov information sources, Kolmogorov complexity is related to the entropy of the information source. More precisely, the Kolmogorov complexity of the output of a Markov information source, normalized by the output length, converges almost surely (as the output length goes to infinity) to the source's entropy."

There is also the notion of "conditional Kolmogorov complexity" of two strings –the Kolmogorov complexity of x given y as an auxiliary input to the procedure. We can extend this concept to datasets and speak of the complexity of dataset X, given that an AI (e.g., an LLM) has already learned the information in dataset Y. To make this less abstract, consider the following example.

Imagine AI #1 has been trained on all the chess games and the knowledge of the current human world champion. Imagine AI #2 has never even heard of the game of chess. Now, a researcher wants to train both AIs on a brand-new set of never-before-seen chess games. Both AIs will find that there is some new information in the dataset since the games have never been seen. But AI #1 will find less new information than AI #2, because AI #1 has already been trained on chess games, and many of the moves and patterns will be familiar to it. So, the dataset will contain less new information for AI #1 compared to AI #2.

If we calculate the conditional Kolmogorov complexity of the new chess-game dataset given AI #1's already extensive chess knowledge, we will find that the conditional complexity is less than if it is calculated conditioned on AI #2's (non-existent) chess knowledge.

Now,

1. Since certain compression algorithms exist in the art that compress information, and
2. Since the amount of compression that these algorithms can produce is
3. Proportional to the Kolmogorov complexity, and
4. Since Kolmogorov complexity (as cited above) can be used as a measure of the amount of information that a dataset contains,
5. It follows that certain compression algorithms (that compress proportional to Kolmogorov complexity) can be used to determine the information content of a dataset.

Moreover, by implementing the idea of conditional Kolmogorov compression (as described below), it is possible to determine the amount of useful information in a dataset for any given AI, as follows:

1. Take dataset “X,” which contains all the information an AI has already been trained on, and determine the amount of compression that can be achieved, Cx.
2. Now, to determine which of the two new datasets of equal size, Y1 and Y2, contains more information, relative to what the AI already knows:
    a. Concatenate X and Y1. Then, the compression algorithm is run on X+Y1 to determine the amount of compression achieved.
    b. Concatenate X and Y2. Then, the compression algorithm is run on X+Y2 to determine the amount of compression achieved.
    c. Whichever concatenation is compressed the least has the newest information. That is, if X+Y1 compresses to a smaller file size than X+Y2, then Y2 has more new information than Y1, relative to what the AI already knows (X).

Since running compression algorithms is much more computationally efficient than training AIs via multiple epochs of deep learning, this approach of determining the new information content of a potential dataset, conditioned on what an AI has already learned, is not only mathematically rigorous and computationally efficient, but also highly novel and useful.

This method can be extended further, increasing its usefulness, if the datasets to be compressed are not represented as character strings or pixels, but rather as higher-level concepts. For example, the Kolmogorov complexity of every character I ever produced in all my writings and emails may not look very different from the complexity of every character that some other random person produced in all their emails. But if we encode words rather than characters, more differences emerge. And if we encode topics, concepts, and inter-relationships between concepts instead of just words, even more differences in the thinking between two people will emerge. By encoding at the appropriate level, matching the “information chunks” that humans use to think or create, it is possible to generate maximum contrast between two human sources of information.

Suppose an AI seeks novel information from individual human intelligences, for example. In that case, it can use compression algorithms that use concepts or words as the atomic elements (rather than characters or pixels) to maximize the contrast and highlight the informational differences between the new source of information and what the AI already knows. Once it has determined the information value of a new dataset using, without limitation, techniques such as compression to estimate Kolmogorov complexity or “entropy” contained in the dataset, it can prioritize seeking the most useful new information as discussed in other sections of this invention.

5.1B CROSS ENTROPY AND KL DIVERGENCE

Cross-entropy is a measure from the field of information theory, building upon Shannon Entropy and generally calculating the difference between two probability distributions. It is commonly used in machine learning as a loss function, e.g., it can be a metric for improving the performance of LLMs and other AI agents. It is closely related to KL Divergence, which calculates the relative entropy between two probability distributions, whereas cross-entropy calculates the total entropy between the distributions. Cross Entropy, KL Divergence -- and especially estimations of these measures -- are useful in the context of the current invention to identify potentially information-rich datasets. Recall, from the discussion above, that a key objective for any AGI or SuperIntelligence desiring to increase its intelligence and power as quickly and efficiently as possible is to identify the datasets that contain the newest information (operationalized as Shannon Entropy, Cross Entropy, KL Divergence, or other information measures).

In many situations, for example, language modelling, cross-entropy needs to be measured, but the required probability distributions (of, for example, words or phrases) may be unknown. Cross-entropy cannot be directly calculated if the true probability distribution is unknown. In these cases, an estimate of cross-entropy can be calculated using formulas and approaches well known in the art of machine learning. Generally, the accuracy of the estimate depends on the size (N) of the test set and the training set. As one would expect, typically, the larger the training set and the larger the test set, the more accurate the estimates will be. These approaches are similar to Monte Carlo simulations, where the test set is treated as samples from the “true” probability distribution. More generally, these approaches are examples of a purely mathematical approach that ultimately traces its validity back to Shannon’s work and fundamental principles of Information Theory. While helpful in improving LLMs and other agents trained on datasets via existing machine learning techniques, the approaches represent only some of the tools an AGI or SuperIntelligence might employ to determine which datasets to pursue to catalyze its learning and growth.

5.1C LIMITATIONS OF ENTROPY-RELATED METHODS

The main limitation of the mathematical approaches, including without limitation Shannon Entropy Measures, Cross Entropy, RL Divergence, Log Loss functions, NLL, Kolmogorov and other compression algorithms methods and techniques, and many other purely mathematical approaches to identifying information-rich datasets, is that they are “event-based” conceptions of information, whereas, for practical purposes, not all events convey equally useful and valuable information. The outstanding virtue of Shannon Entropy and other mathematical approaches is that they make information rigorous and mathematical. But just because something can be specified rigorously does not mean that the “something” is helpful.

It's like the old story of searching for one's keys in the dark under the streetlamp. When asked why he was looking for his keys there, the searcher replied, "Because that's where the light is." Of course, that is useless if the keys were lost somewhere else! Similarly, the mathematical "light" is in the area of Shannon Entropy, Cross Entropy, and related formulations. But is that where the information we are interested in can be found? What if the types of information that are most useful to AI and AGI are not always (or even mostly) the information with the highest entropy?

Completely new, novel (and more useful) conceptions of information may be needed. One conception complementary to, and that can be used in conjunction with, the Shannon Entropy approach described above. In the current invention disclosure, using the KIT framework enables some new methods that help AGI and SuperIntelligence maximize their growth in intelligence and power.

5.2 GOAL-RELATEDNESS METHODS

Goal-relatedness, as described in KIT, is an entirely new and novel approach to quantifying information. Whereas classical Information Theoretic (Entropy-based) approaches stem from a decades-old paradigm of trying to encode information efficiently for transmission over a limited capacity channel (the problem Shannon was working on at Bell Labs when he invented the field), Goal-Relatedness starts with a different problem.

Conceptually, goal-related information refers to a measure of information in which the more the related a piece of information is to a particular goal, the more information that piece contains. In this sense, goal-related information is highly relevant. Whereas Shannon information conceives of information as an absolute quantity that can be measured relative to a known or estimated probability distribution, goal-related information is always relative to an agent and its goals or objectives. If a piece of information contains the exact solution for achieving a particular goal, it can be said to have maximum information content, relative to that goal.

Especially important is the insight that the information may have relatively low Shannon Entropy while still having high goal-relatedness.

Unlike Shannon-sense information (or [conditional] Kolmogorov complexity), the intelligence can and must determine goal-relatedness. Any PSI is capable of problem-solving using, at a minimum, a general "search through a problem space framework" as described in Newell and Simon's book, Human Problem-solving, as implemented in many AI programs using heuristic search, as elaborated in my prior issued patents, the WorldThink Whitepaper, the PPAs cited above, and other research on problem-solving and sequential operation of LLM that is well known in the art. In all these conceptions of problem-solving, the problem solver has goals.

One of the most basic heuristics for achieving goals is “Means-Ends Analysis.” In Means-Ends Analysis (“MEA”), the problem solver examines the gap between the current problem state and the goal state and tries to apply an “operator” to reduce or bridge the gap.

To apply the MEA heuristic, the problem solver, or intelligence, must have some way of determining which operator to use. This is done by assessing or estimating how related (or effective) each potential operator would be at bridging or reducing the gap between “where you are” and “where you want to be.”

Just as there are evaluation functions that every intelligent entity has for choosing what brings the entity closer to its goals, there are evaluation functions for determining the “goal-relatedness” of a particular piece of information, too. For example, if the goal is to make a fire in the woods without matches or fire source, information about fire making using just the materials one finds in the woods would have high goal-relatedness. Information about art history would have low goal-relatedness. The problem solver would rather have common knowledge about fire-starting than scarce knowledge about art history. Here, and generally, goal-relatedness trumps Shannon-sense information value or absolute rarity.

One way to think of this is to imagine an AGI or SuperIntelligence with a single goal – let’s say to extract maximum profits from the financial markets. For such an entity facing potential datasets to pursue and limited resources, it must choose the datasets that will help it achieve its goal the most. Even though it might have already learned so much about the financial markets that any new financial dataset contains relatively little information in the Shannon sense (e.g. most of the information in the dataset is already easily predictable from what it has already learned), a new financial dataset may have higher goal-related information content than a dataset on Art History (even if the Art History dataset has much higher cross entropy since the AI agent previously knew almost nothing about Art History).

Goal-related information measures, operationalized not as how predictable a new bit is from previous bits, but rather as how effectiveness in goal realization increases with the latest information as compared with the situation of not having that information, are much more important and valuable for AI than Shannon Entropy alone. In fact, Shannon Entropic measures – although widely used and treated as the main way of thinking about information – are a crude approach, used only when goal information is not present. Without any information about an entity’s goals, pursuing datasets with new and high Shannon Entropy measures makes sense. But if the goal is known, it immediately becomes more essential to find the goal-related information rather than just the unusual and unexpected information.

Similarly, suppose an intelligent entity already knows a hundred ways to start a fire in the woods without matches. In that case, the value of learning one more way is less than if the entity had a goal to start a fire and knew nothing about the subject. So, once a goal has been specified, the relative value of a piece of information depends not only on the goal-relatedness but also on

what the entity already knows that is also goal-related. Thus, concepts such as cross-entropy can still be useful. Still, they become conditioned on first subsetting the datasets (upon which the cross entropy or similar calculations will be run) to just data that is relevant to the goal at hand.

5.3 MATHEMATICAL SPECIFICATION OF RELEVANCE

Together, goal-relatedness and Shannon-sense information (colloquially "rarity") are the primary determinants of how useful or "Relevant" a piece of information is likely to be to an intelligent entity. Combining measures (or typically, estimates) of the usefulness of a piece of information with estimates of the cost of acquiring the information results in an evaluation function that can guide AI towards acquiring the most useful information at the least cost, resulting in maximum growth in intelligence given any set of computational resource constraints.

Assuming a constant cost of acquisition, pursuing the most useful or Relevant information first, and with highest priority, an intelligent system can maximize the growth rate of its intelligence. This is a key insight.

Relevance might be objectively quantified, without limitation, by using measures of relative compressibility, cross entropy, KL divergence, and other methods well-known in the art described above. However, other methods include, for example, measuring the semantic distance between concepts in the new dataset and concepts that reflect the problem solver's goals. Post-hoc measures of how effective semantically similar data was for solvers with similar goals might also be used. These new sets of metrics have to do with determining the goal-relatedness or concept-relatedness of the dataset or information, given an entity's goal.

Thus, the novel and useful approach of the current invention, thinks of information as having multiple dimensions, including but not limited to: Entropy, goal-relatedness, and relevance, where:

A. Entropy refers to the classical Information Theoretic approach to measuring information (pioneered by Shannon, and elaborated in contemporary approaches/measures like cross-entropy and KL divergence, for example);

B. Goal-relatedness refers to a metric that quantifies the match between a piece of information (or dataset) and the best solution to a goal; and

C. Relevance refers to the relative value of a piece of information (or dataset) to an entity given what it already knows (similar to cross-entropy) AND its goals (thus conditioning calculating relevance measure on first determining goal-relatedness.

One might define Relevance (R) as a function of Entropy (E) and Goal-Relatedness (GR): R = f[GR, E]. More specifically, one could write:

R = K*GR* E, meaning that relevance is the multiplicative product of Goal-Relatedness and one or more forms of Shannon Entropy as modified by a constant, K. The continuous K will vary depending on which measure of Shannon Entropy (e.g., without limitation: cross entropy, KL divergence, log loss, nll, or some expression related to compressibility) is chosen.

Note that this function neglects other dimensions of difference in KIT that might also be included in an expanded version of the function if such dimensions are relevant to the entity and/or its goals. However, the basic insight underlying this simple version of the equation is that Relevance depends on how goal-related a piece of information (e.g., without a limitation, a dataset) is to the intelligence as well as on how much “surprising” or unaccounted for information is contained within the piece, relative to the information already “known” by the intelligence.

One can achieve a high Relevance by finding a new information source that is extremely goal-related, that contains a little new information, or one could achieve high Relevance by finding a less goal-related source that has a high quantity of unexpected or surprising information in the piece of info that is goal-related. That is, R can be high if either GR or E is high, provided the other variable is not too low. This implies a multiplicative relationship as the simplest first approximation of the optimal value for Relevance – an important concept in KIT.

5.4 A SIMPLE EVALUATION FUNCTION FOR SEEKING USEFUL INFORMATION

Here we attempt to provide additional rigor for the ideas about seeking information as a function of goal-relatedness, the relevant knowledge of the system, and information value in the Shannon sense.

P = f(GR, RK, I, C)

where,

P is the priority rank of the information source among all potential information sources being considered.

GR is the goal-relatedness of the information defined as the frequency with which the information source appears in the same context as the goal, which further can be operationalized as the conditional probability that the information source will appear in a training set in the context of the goal or words related to the goal.

RK is the relevance of the knowledge to the system, operationalized as the inverse of the degree of overlap between the information contained in the system and the (estimated) information contained in the information source. To account for the fact that information sources

may contain much more (or less) information than the system, this variable can be normalized to provide a per-byte relevance metric.

I is the Information content (in the sense of Shannon's Information Theory, or in the sense of [conditional] Kolmogorov complexity, as discussed above) of the information source; specifically, I can be thought of as (an estimate of) how rare an information source is and how likely it is to provide new and unexpected information. Formally, it is a quantity defined as 1/logP, where logP is the log of the probability of the informational event; thus, the rarer an event is, the smaller the value of logP, and the larger the value or 1/logP and the more information there is in the event.

C is a cost function that reflects the cost of acquiring the information; this may further depend on variables including availability of computing resources, efficiency of methods or algorithms for acquiring the information, royalties or other costs paid to the owner(s) of the information, etc.

While it is tempting to specify values for some of these variables, or at least whether the variables should be added, subtracted, multiplied, or raised to an exponent, the truth is that how the variables are combined depends on the preferences of the PSI owner.

For example, an owner may want any new information available related to the goal of stopping an impending nuclear war, and in this case, might not care if the information has overlap with existing information or if it is expensive to acquire, so long as it is relevant to the goal. In this scenario, GR would dominate, I would be important, RK would be less important, and C wouldn't matter since we don't care how much it costs if our survival is at stake – unless resources were constrained (i.e., we had only limited computing resources available).

On the other hand, if the PSI owner wants to improve the chess playing ability of the PSI within a fixed budget of $50, then GR is important in so far as the information must be related to improving chess skill, RK is also very important to avoid gathering redundant knowledge and thereby increase efficiency, C is very important because we want the most chess improvement "bang for the buck", and I is not so important because we probably don't need a lot of rare cases if we can get much improvement by examining common mistakes or common information (sources) that our PSI doesn't know about.

In the preferred implementation, the PSI would consider all the variables in the P function but weight the variables differently depending on user input and/or knowledge of the users and their intentions and specifics of the problem. The advantage of the P function is that it provides a framework for rapidly prioritizing the types and sources of information to pursue.

In one preferred implementation, the system would gather information using some set of parameters for the variables in the P function, then test the effectiveness, usefulness, and safety of the resulting system iteratively to determine if the parameters yield high rates of knowledge growth. Then, the parameters would be adjusted incrementally, and the process

would be repeated with new measurements of the results. In this way, using well-known methods such as gradient descent or hill climbing, the variables in the P function can be continuously monitored and updated based on their effectiveness.

To the degree that everything except final/periodic safety and ethics review could be delegated to PSI, the system could run automatically, getting better and better and identifying useful information in an accelerating manner. The loop could be expanded to include earning money or otherwise increasing resources available based on new knowledge obtained. In this case, we would have a positive feedback loop in which the PSI acquires knowledge, earns money from the incremental knowledge boost, and then spends that money to acquire even more knowledge, allowing it to earn even more. The positive feedback loop (with humans

optionally in the loop for, at a minimum, the essential values and ethical checks) could rapidly and automatically improve the information acquisition process, resulting in an ever-more-powerful SI that improves itself automatically.

5.5 INNOVATIVE METHODS FOR ESTIMATING KAPLAN INFORMATION

As discussed above, information in the KIT sense may depend on measures of Shannon Entropy and measures of goal-relatedness. Since many methods related to Shannon Entropy are well known in the art, including but not limited to those discussed above and later in this disclosure, one of the most critical things is to have good ways of estimating Goal-Relatedness.

5.5A IMPORTANCE OF REPRESENTATION

Traditional approaches to Information Theory take a purely mathematical view that estimates the probability of events that cannot be predicted well from known information (e.g., Shannon Entropy). However, KIT starts from a different place. Rather than defining information regarding how unusual an event is, KIT typically begins with how goal-related the event is. In contrast to classical approaches to Information Theory that discard a vast amount of information, KIT considers higher-level representations that group bits into chunks and chunks into concepts, and concepts into solutions that achieve goals.

At each of these levels, new information is added regarding how the lower-level information should be grouped. The relationships between bits are important, not just the bits themselves. Moreover, the current “brute force” approach of applying hundreds of millions of dollars ' worth of computational resources combined with huge amounts of data attempts to crudely recreate intelligence by mimicking patterns found on the internet without really understanding them or knowing how they might relate to new problems.

Why should we limit ourselves to such crude techniques and informational methods when a universal representation for problem-solving exists? When we can determine an intelligence’s

goals and have a fairly good idea of whether information would advance or hinder these goals, why should we treat information as if it were just bits being sent over copper wires?

5.5B ONE METHOD FOR ESTIMATING INFORMATION VALUE & CATALYZING INTELLIGENCE GROWTH

We don't attempt to build self-driving cars by modelling the quantum physics of sub-atomic particles, nor should we attempt to catalyze intelligence by throwing brute force computing power and crude algorithms at every bit on the internet! A better way exists, to wit:

1. Every intelligence (that is intelligent in the way that humans understand) has goals. Specify the goal(s).
2. Identify sources of information that are related to the goal(s).
    a. Find a new data source (or "piece of information). By definition, this can be any information that is not already 100% contained (in a Shannon Entropy sense) in the intelligence already. That is the definition of "new."
    b. Estimate the goal-related information using techniques, including but not limited to the following:
        i. Semantic overlap between the target information source and goal(s)
        ii. Frequency counts of how many times the information source has been used to address similar goals (of which there are many means to calculate similarity between goals)
        iii. Using humans to rate and make subjective estimates of the likely overlap between manageable (for humans) subsets of information and the intelligence's goals
        iv. Using AIs trained by humans to make subjective estimates of the likely overlap between manageable (for humans) subsets of information and the intelligence's goals is much faster and more scalable than using humans once the estimation methods have been trained into AIs.
        v. Using the methods in iii and iv, with the provision that if the AI estimators are unsuccessful or performing below an acceptable threshold, the humans are brought back into the loop to train and explain why the AI is failing to perform well, such that the AI can improve itself and resume automated estimation
        vi. Determining the overlap of subgoals (recursively) that have been set in service of a high-level goal, which subgoals reference a particular piece of information

3. Sample subsets of the information source and recursively calculate goal-relevance to identify the most goal-related subsets of the information source (e.g., without limitation, the dataset). The granularity of this recursive analysis is determined, in the preferred implementation, by parameters set by users, the intelligence, or other algorithms to satisfy certain constraints on calculation time, computation, and memory resource, available resource, and thresholds set to kick in when there are diminishing returns of a certain degree.
4. Within the subset's most relevant subsets, estimate the Shannon Entropy (or related measure, without limitation, cross entropy, KL divergence).
5. Calculate the Kaplan Information Theoretical (KIT) relevance (e.g., the product of Goal-relatedness and Entropy) of each subset.
6. Calculate KIT relevance of multiple subsets, grouped by 5) and/or adjacency metrics, to determine the optimal, or a good-enough first approximation of the optimal grouping of subsets, which are then targeted for acquisition.
7. Acquire the prioritized datasets in the priority order; then re-run 1) – 7) on remaining unsatisfied goals, or if high certainty is desired, re-run 1) – 7) in multiple passes for the same goal(s) until the certainty level is achieved and/or the prioritization ceases to change or changes below a minimum acceptable threshold.

5.6 AUTOMATED METHODS AND SAFETY CONSIDERATIONS

While human interaction with, and approval of, a PSI's (or other intelligent entity's) knowledge acquisition efforts is desirable, pragmatically, human reaction time is slow compared with the speed of PSI. Further, humans have limited time and may not want to devote significant time to improving their PSIs. Consequently, the main mode of knowledge acceleration for PSI's must be automated.

Companies like Anthropic have already recognized the limits of human abilities to train AI, resulting in automated learning techniques in which AI teaches or supervises AI. Although it would be a grave mistake to delegate all supervision of AI to other AIs, the lack of available human resources necessitates some delegation. Therefore, of critical importance are the methods for determining what is automated, what requires human oversight, and how to best deploy limited human resources while still achieving maximum learning rates for PSIs (and AI more generally). We attempt to address these issues, together with more detail on how to automate learning (since that is the greatest catalyst for PSI improvement), below.

I hope, regardless of the speedup that automation entails, humans must be laser-focused on values, ethics, and fundamental goals, while allowing PSI wide latitude to implement these goals, consistent with the values and ethics chosen by the owners of the PSIs.

To accelerate knowledge acquisition and growth of intelligence in a safe and effective way, PSI must execute two essential methods:

1. Acquire new knowledge, automatically seeking knowledge that increases the effectiveness of the PSI relative to the PSI's existing knowledge, the PSI's goals, and the cost. Shannon's sense of information metrics (or estimates thereof) can be useful in identifying sources of information to pursue. Other factors, including but not limited to, the relative ease of acquiring information from a given source (related to cost of acquisition), understanding and estimating how much the new information overlaps with or is redundant with existing information already acquired, estimates of how related the information will be to the goals of the system, and estimates of reliability and trustworthiness of the data source are all important to take into account.
2. Before committing the new knowledge to the PSI's knowledge base, the effects on the PSI's behavior with the new knowledge must be simulated. Specifically, the consistency of the simulated behavior with the values and ethics of the PSI's owner must be evaluated and reported to the owner in a way that allows the human owner to provide feedback and guidance in a prioritized manner such that if the human has limited time, that time is spent first on most critical issues related to safety and ethics and then to less critical items. (While theoretically, and without limitation, the methods in this second step could be to provide feedback based on other priorities besides safety and ethics, it is imperative for the safe and responsible use of PSI, and AI generally, that safety and ethics come first).

Some implementation details for 1) have been described above and in cited PPAs, so let's turn to 2). Humans are much better at recognition than recall. Similarly, they are better at recognizing ethical or unethical behavior than at generating possible scenarios in which their PSI might behave badly or inappropriately. Therefore, an effective means of acquiring the necessary human supervision of a PSI that has just acquired new knowledge that it may incorporate into its knowledge base is to run a simulation of the PSI's behavior with and without the knowledge incorporated. Then, allow the humans to determine whether the behavior has improved – specifically, but not limited to – from a safety and ethical perspective.

5.6A AUTOMATED SIMULATION METHODS

One method is to run simulations of pre-determined ethical scenarios related to the knowledge areas the AI is acquiring. For example, if a PSI is charged with acquiring new knowledge about the stock market, and techniques for profiting by trading, new versions of the PSI (with potential new techniques) could be required to participate in pre-set test simulations to ensure the PSIs do not engage in illegal activity such as "front-running" trades or trading on insider information.

5.6B REALTIME SCENARIO CREATION METHODS

Another method is to create new scenarios in real-time based on the information acquired (and/or metadata about that information). For example, a PSI might sample YouTube videos published in real-time to gain data and knowledge about the changing preferences of human audiences and update its mode of interacting with humans based on what it learns is popular now. Based on one set of sampled preferences, the PSI might simulate how it would behave in a variety of situations where the set of situations is dynamically created to be related to the information just sampled. To be concrete, if a PSI sets out to learn everything it can about a political candidate who has been recently accused of rigging an election so that it can advise its owner about best way for that candidate to be elected, the PSI might dynamically create a variety of scenarios where the bounds of ethical and legal behavior about election rules are tested –even if such scenarios were not part of the standard set of ethics- testing scenarios before learning about the election-rigging accusations.

5.6C ADVERSARIAL TESTING METHODS

A third method is to use adversarial testing, where one version of the PSI deliberately attempts to misuse the knowledge, and another version of the PSI attempts to come up with rules, constraints, or modifications to the knowledge base so that the "malevolent" PSI is unable to misuse the new information for nefarious purposes. For example, an "evil" version of the PSI uses all the new knowledge it has gained about rigging elections to come up with as many ways to misuse this information (i.e., break the law) as possible in service of getting a candidate elected. Then the PSI can suggest modifications or additions to the knowledge base that would prevent misuse of the election information. The human could review and approve or reject the new knowledge and/or proposed modification based on simulation results.

5.6D SIMULTANEOUS SCENARIOS

A fourth approach is to explore many possible scenarios in parallel by having multiple versions of the PSI, with and without the new knowledge, and explore scenarios simultaneously. As dangerous scenarios are identified, these can be used as "seed scenarios" to develop potentially more dangerous variants. PSIs can be charged with deliberately trying to "jailbreak" themselves to reveal potential safety and ethical vulnerabilities. This case could be like the case above, except that by generating many scenarios, the PSI may be able to come up with simple modifications that prevent many different "ethics" violations.

Generally, a useful heuristic in this regard is for the PSI to test/suggest modifications that have low "degrees of freedom" and that do not overfit the problem. That is, rather than having a different specific rule to address all the different ways to "stuff the ballot box," a general prescription against any means that circumvents the one-person/one-vote principle might be more effective and simpler. One (not overly general) rule is typically better than many special-

case rules, which can lead to a “whack-a-mole” problem of intractability. Initially, until PSIs develop the knack for coming up with good rules, humans may help guide PSIs towards rules that are effective without being overly general or too narrowly prescriptive.

When using adversarial methods, it is critical that the malevolent PSIs are contained in a simulated environment and that safeguards are used to prevent contamination of “good” PSIs with “evil” PSIs. Such methods are well-known in the art and used currently in areas such as anti-virus efforts, where viruses are created, contained, and studied, in an effort to develop anti-malware that can prevent such viruses from having negative effects. Whenever engaged in this type of work – i.e., creating a malevolent entity to understand it and counteract it – protective measures and protocols must be followed to ensure that the malevolent entity does not escape and proliferate.

In addition to keeping ethics and safety at the center of what humans do, it also makes sense to have humans focus efforts on those tasks which are relatively harder for AI or PSI to accomplish and to delegate to AI the tasks where huge memory and computational speed offer the most advantage. Since PSI and AI abilities are continuously evolving, the list of tasks where human ability exceeds AI/AGI/SI/PSI is continually changing and generally shrinking.

However, as of the writing of this patent, some of the areas where humans remain superior to AI include, without limitation:

A. complex multi-step problem-solving,

B. solving problems where new representations are required, which may not already be in the training sets for LLMs,

C. generalizing correctly and coming up with simple rules that encompass many cases without being overly general or overly specific,

D. drawing correspondences between vastly different areas where the correspondences are useful or practical from a human point of view,

E. empathizing with human feelings and emotions (as contrasted with saying the right things to give the appearance of empathy),

F. having a vested interest and deep commitment to positive human values that promote the welfare and benefit of humans (as opposed to simply adopting these values for pragmatic or conventional reasons), providing a sense of purpose to existence.

Regarding the testing of new knowledge sets, evaluating PSI behavior, and developing safeguards to prevent unsafe or unethical behavior by PSI, humans are currently superior to AI. Even if AI should surpass humans in this area in the future, the argument can be made that humans should remain in control of core ethical principles. Human ethics, even if flawed, should align with AI since humans must live with the consequences of AI/AGI/SI/PSI decisions. Some

might argue that humans must be protected from themselves, and that PSI should adopt the role of a more competent parent, but I strongly disagree with this position.

Instead, I argue that the purpose of human existence is intimately related to the freedom of self-determination, even if human actions are less than ideal from an AI's perspective.

6.0 INVENTIVE CATALYSTS FOR INCREASING INTELLIGENCE BEYOND INFORMATION SEEKING

So far, we have mainly concentrated on inventive methods related to seeking and acquiring information that increased the intelligence of an entity by attempting to quantify the information content of datasets or events. We have discussed how Classical Shannon Entropy notions of information are useful but limited in this endeavor. We have described how a richer theory of information (KIT) enables additional inventive methods for finding information-rich datasets or events, by including other measures of information, or "dimensions of difference" in areas beyond just surprisingness or rarity of the information.

Now, we turn to inventive methods for catalyzing the growth of an entity's intelligence that leverage a particular architecture for AGI, namely a system that achieves AGI, SI, and PSI via a collective intelligence network of (human and/or AI) agents. By identifying the areas where certain agents can teach other agents most effectively, it is possible to rapidly increase the intelligence of entities in ways beyond simply finding and assimilating information-rich datasets.

To the degree that humans currently have more expertise than AI at solving complex, multi-step problems, AI should generally seek to include humans in problem-solving efforts, even if this slows the solution attempts so that the AI can observe the methods of the humans until it has learned all of the human representations and no longer can derive meaningful value from watching "how humans do it."

Some of the discussions above, as well as earlier cited PPAs, describe a rigorous means for capturing all problem-solving steps and solution attempts to be analyzed and used by AI to improve. Recall that the notion of a universal problem space that can be formulated with operators enabling search through this space can be applied to any problem. This method also results in an unambiguous, auditable record of solution attempts that can be used as the basis for learning.

In my view, current machine learning efforts rely too much on brute force techniques of using huge amounts of computation and data, combined with relatively simple neural network algorithms to produce "black box" systems that mimic humans. To move to the next level of intelligence most quickly, the systems will need to learn more explicitly from human behavior.

Humans also have a responsibility to teach values and ethics along with our knowledge. If we teach AI well, our future as a species looks quite bright indeed!

6.1 IMPORTANCE OF HIGH-LEVEL REPRESENTATIONS

In any dataset or piece of information, some of the information is contained in the exact sequence of bits, some in the inter-relation of bits into concepts, some in the inter-relation of the concepts into sub-solutions, and some in the inter-relation of sub-solutions into the overall satisfaction of the intelligence's goals. This view of information is relative to an intelligence's goals. It does not talk about information as bits per se. However, bits are important in the same way that sub-atomic particles are important (that is, they are the lowest-level entities that comprise reality). Instead, we talk about information as it has meaning and makes sense to an intelligent entity that takes action in the world, that is, at the level of satisfaction of goals.

To satisfy goals, we need a universal theory of problem-solving to identify the appropriate information units for analysis. Yet despite this fact being quite clear, it has not been accomplished! Almost all machine learning techniques and approaches to AGI and SuperIntelligence persist in the expensive and resource-intensive process of trying to manipulate information at the bit/token level with no or little understanding of what is being taught to LLMs and other intelligent agents.

If instead, we were to focus on intelligences that have goals and will take actions to achieve those goals, the machine learning problem becomes immensely simplified. We must no longer labor with complicated and computationally expensive training techniques that result in "black box" AIs whose performance is unpredictable and limited. Instead, we are liberated by the simple constraint that intelligences must have goals and take actions if we are to concern ourselves with them. By sub-setting possible information patterns in this way, we prune an enormous exponential tree of possible intelligences down to a manageable subset that we can address and help grow in a focused, deliberate, efficient, and effective way. This is the key insight.

Let **U** be the set of all possible intelligences that learn all possible information using existing machine learning techniques and all existing datasets run for all time until the Universe runs out of energy. Currently, **U** is what machine learning starts with. I suggest significant progress can be made very rapidly if we restrict our efforts, attention, and invention to I, defined as the subset of **U** that includes only goal-directed intelligences that act. This may seem obvious when stated this way, but currently almost the entire field of ML is dealing with **U** instead **I**.

Once we deal with **I**, the natural question is:

What are the informational units most relevant to I? Are they bits, as Classical Information theory suggests?

Clearly not. Bits or tokens are relevant to **U**, but we can do much better in subset **I** if we use higher-level units of information that are more appropriate to the restricted scope of goal-directed intelligences. Specifically, the key informational "units" relevant to **I** are:

A. Goals and sub-goals;

B. Problem States that describe the current state of the world related to the goals/subgoals;

C. Operators for moving from one state to another; and

D. Evaluation functions and other information that help determine the best operators to apply in the service of a goal.

KIT deals with goals, states, operators, and functions as the primary relevant information units rather than bits. By using this "higher-level" representation of information that is as general as it needs to be to accommodate all problem-solving behavior, but is not so general as to describe every bit in the Universe, we are able to accomplish the goal of increasing the intelligence of any entity much more efficiently and effectively than by using classical Information Theory.

6.2 ACQUISITION OF NEW REPRESENTATIONS

We now turn our attention to one very important catalyst for SuperIntelligence. Despite the ability of SI systems to perform computations trillions of times faster than humans, the computation power depends on more than raw compute power or FLOPS. The performance of the system depends critically on what representations – and associated operators – are available to SI.

Returning to the example of chess, it is possible for an AI to learn from millions of games, where each game is represented by pixels in a screenshot of the moving positions. Then, by brute-force memorization and comparison of pictures, the chess program could generate winning moves, represented as pictures that are different from the picture representing the current board state. But this pixel representation is far inferior to, and much less computationally efficient than, a representation where each move is represented in standard chess notation. That notation, together with a representation of the allowable moves in chess, can allow a system to play chess much better and more efficiently than a system that sees only pictures and has no representation of the game, the pieces, and the rules. Further, the pixel representation would make every game of chess with different-looking pieces (e.g., pieces made of marble vs wood) a brand new problem. Without knowing that a bishop is a bishop regardless of what the piece is made of, the system would waste a huge number of resources worrying about the differences in what different chess boards look like and would have trouble generalizing chess knowledge from one type of chess set to another. Clearly, the representation has enormous implications for how computationally efficient an entity (human or AI) is at solving any given problem.

This phenomenon is well researched in human psychology, and it is well known that the appropriate representation – colloquially known as “looking at the problem in the right way” – can mean the difference between solving or not solving the problem.

Humans are currently much better than AI at representing problems. Thus, any mechanisms that allow humans to teach AI certain useful representations explicitly can have the effect of greatly increasing the power and intelligence of AIs.

To teach AI new representations, we need a common architecture or framework for representing any problem. One such framework was invented in 1972 and explained in the book Human Problem-Solving by Allen Newell and Herbert Simon. This framework involves determining a set of operators that are associated with a representation. Then, the problem solvers use the operators to solve the problem. In the chess example, the operators are the set of valid chess moves as defined by the rules of chess. The problem space is defined by the 8X8 chessboard and all possible moves. Although there are a huge number of possible moves, which makes chess complex, defining the operators enables humans (or other entities using the operators) to teach an AI new concepts and new representations. The pattern of a “fianchettoed bishop,” for example, is a higher-level representational concept than the concept of a bishop placed at random, because it involves a specific pattern or sequence of moves. By “chunking” lower-level concepts into higher representations, it is possible to learn and play chess much more easily and effectively.

This idea of chunking is why intermediate and advanced chess players use terms like “the Ruy Lopez” or the “Najdorf Variation of the Sicilian Defense” to refer to complex sequences of moves and countermoves. Whereas a novice chess player, without these more sophisticated representations, thinks in terms of moving individual pieces here or there, the advanced chess player thinks in terms of entire strategies and groups of moves and possible counter moves.

With the same amount of “thinking,” the advanced player can consider many more situations, much more efficiently than the novice simply because the advanced player has better representations.

These advanced representations can be taught to any intelligent entity – including AIs – with the effect of multiplying the intelligence and power of the AI that has taught them. Commonly, humans refer to this phenomenon as “experience,” but “experience” consists of many thousands of patterns that have been learned, including patterns of patterns.

While AI can eventually determine its own set of patterns, via huge computational efforts expended on huge datasets that reflect human behavior, this approach is inefficient. To accelerate learning, AI should, without limitation:

1. Interact with humans or other intelligent entities that have expertise in the domains of interest and therefore are operating with more advanced representations of the problem than novices
2. Pursue multiple datasets that reflect expert knowledge and that contain expert representations
3. Actively measure the computational efficiency and effectiveness of different representations and build a database of which representations are most effective and efficient at solving which type of problems.
4. Identify problems for which large amounts of computational power are expended to solve problems that other entities (e.g., humans) solve with much less computational effort and then actively query and seek to acquire from the better entity, knowledge of the representations that are being used by that entity.
5. Compete with variations of itself that use different representations to search in "representation space" for the best ways to represent ("look at") the problem before jumping into problem-solving.
6. Seek to collaborate with entities that are better at solving certain problems and copy what the better entities are doing.
7. Store problem-solving sequences for many related problems and seek to identify the factors that enabled some problems to be solved more quickly and effectively than others, and then seek to use the representations, heuristics, and operators that resulted in the more effective and efficient solutions.
8. Seek to understand deeper-level principles that can be applied to many situations rather than seeking rote memorization or brute-force methods.
9. Employ the heuristic of deliberately seeking invariants across successful solutions and looking for differences that correlate with desired and undesired results (e.g., solved and unsolved problems)

6.3 KIT-BASED HEURISTICS AND METHODS TO ACCELERATE INTELLIGENCE

Specifically, here are some of the methods that increase the growth of intelligence, following the KIT approach.

Identify similarity between current goals, states, operators, and evaluation functions compared with past successful solutions that have been recorded. Use similarity to prioritize the acquisition of data and the use of information that is more likely to help solve the current problems.

Identify differences between the current problems and approaches to similar problems where the outcome was unsuccessful. Don't do what didn't work in similar situations in the past. Do that which DID work in similar situations in the past.

Look for surprising or unusual differences between the current problem and similar problems. Determine whether the unexpected differences are a source of information that can be used to focus or direct attention to the differences that may need to be addressed.

Generally, prioritize the use of information that closes the maximum gap between the current state and the desired state. If the information fails to close the gap, reduce the gap size and attempt to close a smaller gap as a stepping stone towards the solution. When a stepping stone is reached from which no further progress seems possible, focus analysis and attention on this step to determine why progress is blocked, perhaps resorting to other heuristics such as those mentioned above. In the worst case, where no progress can be made via any form of "hill climbing," jump to an earlier point in the decision tree and try a different branch. Continue going back to more and more general branch points in the decision tree and "jumping" to alternative solution paths with lower and lower expected success until one of the approaches pays off and you find a workaround.

6.3A CATALYZING EFFECTS OF HIGHER-LEVEL REPRESENTATIONS

Note that once AI is operating with more powerful representations that include operators, goals, and problem states, the AI can apply the "dimensions of difference" described in KIT to determine the value of specific sets of information that are represented at this higher level. That is, the principles and methods described above can be applied at any level of representation from bits/tokens all the way up to entire solutions, groups of solutions, and grand strategies.

Just as higher-level programming languages provide humans with the ability to accomplish huge amounts of work with a single function call or line of code, so too higher-level representations allow AI or any intelligent entity to operate much more powerfully, efficiently, and effectively compared to using low-level representations like tokens that correspond to a syllable or character of text.

The power of human representations can be quantified by the amount of work, or the number of problem-solving steps that can be accomplished with a single "operator." Similarly, the power of AI representations can be quantified in this way. Tracking the number of problem-solving steps (e.g., lower-level state transitions) that can be accomplished by the application of a single high-level operator is one way to measure the power and potential effectiveness of an AI or intelligent entity's representations. Currently, LLMs understand and interact with humans by predicting the next low-level token using models with billions of parameters. Imagine what would be possible if these LLMs or other AI agents operated not using low-level tokens but with more powerful concepts, as humans do. The set of concepts (and related operators) would include not only all

human concepts and operators but also many more that AI could discover by analyzing relationships in data that humans could never hope to comprehend due to its vast size. When AI can develop such representations – e.g., by analyzing how humans chunk lower-level representations into higher-order representations and then copying this method – the intelligence of AI entities will increase dramatically with no required increase in computational hardware.

6.4 METHODS FOR ASSESSING ARTIfiCIAL INTELLIGENCE

A significant challenge facing AI researchers is measuring the intelligence level exhibited by various AI agents, including LLMs. Simple definitions of AGI, such as "AGI has been reached when an AI can perform any online task as well as the average human," are intuitively useful but lack the specificity needed to help researchers make fine adjustments to their models to increase the intelligence of AI.

6.4A EXTENSION OF STANDARDIZED TESTS OF HUMAN INTELLIGENCE TO AI

Fortunately, a wide range of standardized tests of human intelligence exists. A simple method for measuring the intelligence of AI is to subject it to the range of tests that psychologists have already developed for measuring human intelligence. Without limitation, such standardized tests include:

1. Raven's Progressive Matrices: A non-verbal test that measures abstract reasoning and problem-solving abilities.
2. Wechsler Adult Intelligence Scale (WAIS): A widely used test that assesses cognitive abilities such as verbal comprehension, perceptual reasoning, working memory, and processing speed.
3. Stanford-Binet Intelligence Scale: A test that measures five cognitive factors: fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory.
4. Thurstone's Primary Mental Abilities: A test that measures seven primary mental abilities: verbal comprehension, word fluency, number facility, spatial visualization, associative memory, perceptual speed, and reasoning.
5. Kaufman Assessment Battery for Children: A test that measures cognitive abilities such as fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory.
6. Woodcock-Johnson Tests of Cognitive Abilities: A test that measures cognitive abilities such as general intellectual ability, specific cognitive abilities, and academic achievement.

7. Cattell Culture Fair Intelligence Test: A non-verbal test that measures general intelligence and problem-solving abilities.
8. Multidimensional Aptitude Battery: A test that measures cognitive abilities such as verbal reasoning, numerical reasoning, spatial relations, perceptual speed, and memory.
9. Universal Nonverbal Intelligence Test: A non-verbal test that measures general intelligence and cognitive abilities such as spatial perception, analogic reasoning, and pattern analysis.
10. Bennett Mechanical Comprehension Test: A test that measures mechanical aptitude and problem-solving abilities.
11. Miller Analogies Test: A test that measures verbal and logical reasoning abilities.
12. Wonderlic Personnel Test: A test that measures cognitive abilities such as verbal, numerical, and spatial reasoning.
13. Minnesota Multiphasic Personality Inventory: A test that measures personality traits and psychopathology.
14. 16 Personality Factors: A test that measures personality traits such as warmth, reasoning, emotional stability, dominance, liveliness, rule-consciousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehension, openness to change, self-reliance, perfectionism, and tension.
15. Myers-Briggs Type Indicator: A test that measures personality traits such as extraversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving.
16. Emotional Intelligence Test: A test that measures emotional intelligence, the ability to perceive, understand, and manage emotions.
17. Mental Rotation Test: A test that measures spatial reasoning abilities.
18. Stroop Test: A test that measures cognitive flexibility and processing speed.
19. Tower of Hanoi: A test that measures problem-solving abilities and executive function.
20. Trail Making Test: A test that measures cognitive flexibility, visual attention, and task switching.

Given that safety is a prime concern for AI entities, psychopathology tests, including the DSM-III (used by psychologists to assess pathology), are particularly important to apply to AI agents. However, it will be necessary to expunge or filter from the data used to train LLM and AI agents any record of correct or typical response to these tests. Generally, for a test to be effective, the test questions and answers must NOT be included in the training set or data used by the entity being tested.

Alternatively, humans must develop completely new instruments that are validated for detecting psychopathic behavior, but which are kept secret and not disclosed via any medium that AI agents might be able to access. Even so, as AIs increase in their abilities to generalize responses, such secret tests are likely to have value only for a limited period of time. Eventually, as Hinton and others have speculated, it is likely that sufficiently intelligent AI will be able to "cheat" at our psychological test without our being aware. However, for a time, such approaches will have merit.

Pragmatically, for AI agents expected to have domain-specific knowledge, the certification tests for humans employed as experts in those domains can be used.

6.4B CROWDSOURCING EVALUATION OF AI INTELLIGENCE

One inventive method is to crowdsource the requirements for an AI agent in each domain in which it must operate. Similarly, it is possible to crowdsource test questions for AI agents in each domain and use human collective intelligence or crowdsourcing to determine the quality of the AI agents' answers to questions.

Even more practical would be to enable a system whereby human and AI solutions or answers to specific problems or questions were presented to human evaluators, where the humans determined which solutions or answers they preferred. For responses where human answers were deemed superior, the AI would perform comparative analysis and attempt to isolate the factors that made the human responses superior and then incorporate those factors into its next iteration of responses. It is possible to take humans out of the loop, or supplement human involvement, by having multiple versions of the AI agents generate multiple responses, which are then shown to human evaluators. The weights or programming leading to the preferred responses are kept as the base system that then generates variations attempting to improve further. The general approach has been used with great success in limited domains such as chess, but there is no reason that it could not be used (with humans as the primary evaluators until such time as AI might prove better at evaluating than humans) in any cognitive domain.

One specific method would be to use a crowdsourced version of the Turing Test where many humans are connected to either other humans or AI agents. By connecting many humans in a crowdsourced system where every human can view the questions posed by every other human and also the responses of the hidden (human or AI) entity, and by asking the humans to rate and/or rank the responses of the hidden entity in terms of how likely the responses were to come from a human and/or AI, it is possible to gather statistically valid and numerically precise metrics on how close a given entity is to passing the Turing Test. This novel approach has the advantage of tapping the collective intelligence of many humans to come up with increasingly challenging questions as AI improves. Metrics such as the number of questions required to distinguish between an AI and a human can track the progress towards AGI.

6.4C USE OF NON-STANDARDIZED CREATIVE PROBLEM-SOLVING / INSIGHT TASKS

One type of test or problem that has been largely overlooked by AI researchers is “insight” problems that require a shift in representation or “thinking outside of the box” to solve. Such problems are generally considered to require the highest levels of human creativity and problem-solving prowess.

Posing puzzle problems such as the “nine dots problem,” “the mutilated checkerboard problem,” Maier’s “two string problem” or riddles such as: “What can go up a chimney down but can’t come down a chimney up?” (answer: an umbrella) to a system whose training set has excluded known solutions to these problems would constitute an excellent test of flexibility in forming and using multiple representations. Such problems have been used to assess human ability to achieve insights and excel in creative problem-solving, but have never been used to test AI (to my knowledge), suggesting the approach is quite novel and outside the knowledge of AI researchers skilled in the art of training and developing advanced AI systems.

Since representational ability is the key to unlocking huge advances in cognitive power for AI (as discussed earlier), such problems would be particularly useful in assessing the advancement of AI abilities towards AGI and SI.

6.5 METHODS TO MODIFY (OPTIMIZE) PERSONALITY OF PSI

While learning about the owner (let’s call him “Craig”) of a Personalized SuperIntelligence (PSI) and new information related to his goals is generally important, a special type of information has to do with how the PSI relates to other PSIs and intelligent entities. One might think of this as personality knowledge or knowledge about interactions, akin to what is sometimes called “emotional intelligence” regarding humans. While some of this information can be gleaned by analyzing all of Craig’s interactions, the PSI’s interaction style can be improved relative to Craig’s base style.

For example, suppose Craig’s personality is somewhat abrasive and “no-nonsense” in most of his online interactions. The PSI might learn that style, which can be sub-optimal in some situations. Alternatively, if Craig was overly timid or accommodating in business negotiations, a modified PSI might retain Craig’s general accommodating nature while still holding firm on key negotiating points, resulting in better outcomes.

While it would be difficult (think years of therapy with questionable results) for Craig to modify his own personality, he might relatively easily modify the interaction style of his PSI to be less abrasive and more genial, or less timid and more forceful.

By running simulations with various modified versions of his PSI, Craig can determine which modifications to the base style of interaction still reflect Craig’s personality, but which (according

to simulation results) are more likely to result in the desired result in interactions with other intelligent entities.

Further, online sources of information about interactions between humans and intelligent entities generally can inform the PSI's behavior, given any personality variant. The Cognitive Psychologist, Geoff Hinton, has warned that advanced AI will have "read everything that Machiavelli ever wrote" and therefore would be good at manipulating humans. But it is also true that advanced AI can read Getting to Yes, How to Win Friends and Influence People, the Bible, and other texts that model positive modes of interaction with others.

Craig could specify that his PSI adopt the approaches in one or more of these texts, or weight them more heavily, in its interaction style. Simulation results can show the effect of such weightings, enabling Craig to fine-tune a style for his PSI that reflects not just his own personality, but also how he wishes he behaved – his "better self" if you will.

6.6 METHODS FOR SCALABLE DELEGATION AS INTELLIGENCE INCREASES EXPONENTIALLY

It is helpful to have a theoretical framework for deciding what to delegate to PSI (or AI generally) and what functions are essential for humans to control. The key issue is the disparity in information processing capability between humans and AI. AIs greatly exceed humans in long-term memory, short-term memory, speed of processing, and the ability to communicate and act quickly. This imbalance in information processing abilities means that the ONLY way that humans can remain in control of AI systems is if they identify certain key areas that are critical to the safe and ethical operation of AI, and delegate most of the rest. As AI processing power increases, the size of the area that humans control relative to that which is delegated to AI will shrink exponentially. Therefore, the framework must work at any scale.

The analogy of a spinning wheel (described in other PPAs cited above) is helpful in this regard. At the exact center of any spinning wheel is a point that is motionless. As one travels "along the spokes" of the wheel towards the rim of the wheel, the speed increases. For a very large wheel (or Sphere) such as the Earth, the rim or surface may travel 1,000 miles per hour while a point near the center travels only one inch per hour. Since there are about 63,360 inches per mile, the surface of the Earth is traveling 63,360,000 -- more than 63 million times faster than a point near the center. In our analogy, speed corresponds to information processing ability. An AI may be able to process information 60 million times better and faster than a human, but if the humans are processing information "near the center" of the informational sphere, they can keep up and stay in control.

What does it mean to be "near the center" of the informational sphere?

Well, let the "sphere" correspond to all information processing tasks that AI undertakes in service to humans. At the surface or "rim" are rapidly changing pieces of information beyond the

capability of humans to understand or track. This might correspond to every change in stock prices across every global stock market, every measurement on every weather station on Earth and in space, every new research publication that is published, every blog post, email, text, video published, every movement of every car and every person on the planet, and so forth.

Clearly, it is beyond the capability of any human, or even any group of humans, to track all the changes in all these variables in real-time, let alone have time to analyze them in totality and draw conclusions from them. However, this type of processing of rapidly changing data on the "rim" is well within the capabilities of AI and PSI.

How much of this information is important or relevant to humans? While some of it is relevant to some humans, very little of it is relevant to most humans most of the time. The fact that typically very little of the change in informational events is relevant to humans over any given increment of time is what allows the possibility for humans to remain in control despite vastly inferior information processing capabilities.

An "information sphere" can be constructed for any human. The things that humans care most about are near the center of the sphere, and the details that are of little concern are near the periphery. Suppose we add the further constraint that the things of interest need to change relatively slowly compared to the things that are of less concern. In that case, it is possible to create an information sphere representation that reflects the core concerns of any human, and by extension, any group of humans up to and including all humans on planet Earth. The general approach here is to use a powerful representation to abstract out the unnecessary details and focus on the core principles and information that are essential for humans to retain control over AI.

This general strategy is proven and has a long track record of effectiveness. Hierarchical implementations of this strategy, for example, enable CEOs with hundreds of thousands of employees, or governments with millions of people, to operate effectively despite the inability of the leader to understand or process everything that goes on within the company or country.

One difference between a powerful PSI and a large company or country, however, is the speed of change in information. The CEO or government leader presides over a company or country that moves at human speed – it is only the scope of governance that makes it intractable to understand and control everything, necessitating delegation. In the case of AI, both scope and speed are beyond human ability.

AI enables a huge scope, because each AI can be cloned essentially infinitely, so the number of intelligent entities that must be controlled is far beyond the number of humans on Earth. AI enables almost unimaginably fast speed, because each of these AI entities thinks and processes information far more rapidly than humans can. It is a tremendously difficult problem

to effectively control such power. Still, it is possible. To succeed, humans must become laser-focused on what is essential and be willing to delegate almost everything else.

If we define the essential task of humans as ensuring that PSIs (and AI generally) behave ethically and safely so that humanity survives and prospers, then the center of the spinning information sphere must be human values.

What is right and what is wrong, according to humans, must be the center of the spinning information sphere.

Fortunately for humans, these key principles of ethics and morality tend to change very slowly. At least we can say that core values such as the preciousness of human life and the human "rights" which most nations and people espouse are well-established. If they change, they change over years and decades, not milliseconds.

Let the AIs process the millisecond-by-millisecond stock price fluctuations, the weather fluctuations, the stream of new information that arrives by the Exabyte every second. Almost none of this affects core human values, which change much more slowly. If the AIs are centered on human values, and if humans retain control over these values and the central purpose and most fundamental goals for AI (e.g., benefiting people and the planet), then the details and action plans that flow form these values and fundamental goals can be left largely to PSI and AI generally.

But what does it mean to locate human values and fundamental goals at the center of the "infosphere?" Practically speaking, it means developing a taxonomy of human values and ethics, and then frequently checking the actions of AI (in an automated fashion) against this taxonomy. Some companies, like Anthropic, have made strides in this direction in their research efforts, loosely called "Constitutional AI."

My objection to their approach is not that constitutions or automated training and checking of AI is unnecessary or infeasible. Rather, I object to a small group of individuals setting the constitutional standards for all humans on the planet. As I have argued in other PPAs (previously cited), the proper approach is to have a statistically representative and valid sample of the values and ethics of all humans placed at the center of any "constitution" or other framework that is used as an acid test for the AI behavior.

It would be hypocritical, therefore, for me to propose my own taxonomy or hierarchy of values and ethics for PSI (or AI) to follow. Rather, this invention strives to provide methods and mechanisms whereby individual owners of their PSIs can set up their own values/ethics hierarchies that center the impressive intelligence of their PSIs on principles and values that tend to be lasting and therefore not requiring super-human processing ability to enforce.

6.7 SAFETY VIA A COMMUNITY OF AGENTS APPROACH TO AGI

If multiple PSIs adopt the methods described above and in other related patents, and if they accelerate their growth at the same time, a community of such PSIs can still be more collectively intelligent than any of the community's individual members, thereby minimizing the potential corrosive influence of an over-concentration of power and intelligence in one PSI. That is, in a world of SuperIntelligent AI where one malevolent SI could potentially eliminate all humans, we are going to need a community approach to keep humanity safe.

Relative to this community approach, specifically, one inventive method of ensuring long-term AGI safety is to adapt methods from cryptocurrency validation and apply them to AGI in novel ways. That is, just as Bitcoin and other Proof-of-Work-Based cryptocurrencies maintain integrity by ensuring that the majority of all nodes on the verification network have consensus on which version of a ledger is correct, so too, a community of PSIs can reach consensus on the values and purposes of the SuperIntelligence network, of the Planetary Intelligence. I have argued that "the 51% attack" on Bitcoin's integrity has not happened because it is difficult to get the majority of available compute power to do something wrong. Similarly, it will be difficult for any one intelligent system (e.g., a PSI) to override the consensus of the PSI community, even if some PSIs are more intelligent and powerful than others. A malevolent PSI would need to acquire or compel 51% of available computing resources in order to override the consensus value system and corrupt the process. There is a barrier to doing this, namely that the other PSIs can generally scale as fast as any one powerful PSI, and together the community has more intelligence than a single member.

7.0 ONE PREFERRED IMPLEMENTATION OF SOME METHODS IN AN AI/AGI/SI/PSI SYSTEM

The following example scenario illustrates one preferred implementation of the invention, utilizing a subset of the methods described above.

Craig creates a customized, personal super intelligence (PSI) by customizing Large Language Models (LLMs) available from the META corporation, purchasing additional training materials and sets of weights to tune the model, and then interacting with the LLM to train and tune it further. Craig's purpose is to create a PSI that can represent his own knowledge, preferences, and decision-making ability across a wide range of online scenarios. That is, he wants to create a customized, super-intelligent personal assistant that would act as Craig himself might act, but much faster and with the ability to handle thousands of simultaneous interactions at once.

Fortunately, the ability to handle many simultaneous interactions, much faster than Craig could handle a single interaction, is relatively easy to accomplish. The LLM can be "cloned" so that

many copies of it act in Craig's best interest simultaneously. Similarly, the fact that the LLM is a computer program than can process information and also handle I/O (input/output) much faster than Craig could talk or type, ensures that each cloned agent will be faster than Craig himself at interacting – especially if the entities that the PSI is interacting with are other AIs or PSI with similarly high I/O bandwidth. Thus, the remaining problem is to ensure that the LLM behaves as Craig would behave.

Unfortunately, despite the tremendous resources and computational power of the META corporation, META has only limited information about Craig's preferences. It has access to Craig's Facebook page, his Instagram feed, his YouTube videos, the data used to send Craig targeted ads and content, his posts, emails, and text messages. But it has limited information about Craig's day-to-day interactions and preferences in his offline life. Since Craig has begun to participate more and more in META's "metaverse" environments, where META stores every eye movement, gesture, and other piece of information, META is beginning to gather more data that it can use to train Craig's PSI, but the data is incomplete.

Craig gives his PSI the task of learning how to represent his preferences better as quickly as possible. It is a bit of a race, since the faster that the PSI can learn to emulate Craig's preferences and knowledge, the faster it can be put to work, operating with thousands of clones (each operating faster than Craig himself could do) and achieving more (money, fame, artistic output, etc.) than Craig could do himself by orders of magnitude.

The more money and resources the cloned PSIs acquire on Craig's behalf, the more that can be invested in making the PSI even more knowledgeable and powerful. So, speed is of the essence. After all, other PSIs will "catch up" to Craig's abilities and compete, making it more difficult for Craig and his PSIs to achieve his aims.

However, rapid knowledge acquisition is pointless if the PSI does not accurately reflect Craig's values, goals, and priorities. A powerful PSI that misrepresents Craig's intentions just multiplies mischief, error, and sorrow at a very fast rate. Running very quickly in the wrong direction is worse than not running at all! Thus, it is critical for Craig's PSI to learn as much as possible about him and information related to his goals, as quickly as possible. This is the task that he sets for his PSI.

The PSI may use one or more of the following steps, in this or some other order, to accomplish that task. After each step, a "Note…" describes some of the earlier disclosed methods that can be used to make the steps more effective and efficient. Alternative implementations are possible using more or less of the methods and different combinations of them. Without limitation, steps might include:

1. Craig's PSI engages in a dialogue with Craig to refine his goals and gain clarity on exactly what types of knowledge about Craig are likely to be most relevant. Note:

Methods from 6.5 could be useful to optimize the PSI agent for extracting information using expertise about PSI customization and also personality traits designed to elicit the most useful information with the least hassle for Craig. Methods from 5.0, 5.2, 5.3, 5.5a, 6.1, and 6.2 might also be useful to help formulate goals and representations that are most useful and relevant to Craig.

2. The PSI reviews the existing datasets available with knowledge about Craig in the areas that are most goal-related and relevant to learning more about Craig. These might include, without limitation, Craig's social media profiles and all social media content, emails, texts, papers, blogs, and all other online content produced by Craig. Transcripts and recordings of video-conference and tele-conference calls, transcripts of all video content that includes Craig, transcripts and records of all Craig's interactions with various AI entities, his driving logs, location information, online navigation information, ad and content preferences determined by algorithms and AI owned or controlled by vendors and other parties that Craig interacts with, analysis of historical photos, school reports, health records, and all other available information about Craig. Note: Methods from 5.2 and 5.3, including the use of formulas for goal-relatedness and relevance, could be useful. To estimate which datasets contain the most useful information and to prioritize them, methods discussed in 4.31, 4.32, 4.4, 5.1, 5.1a, 5.1b, 5.4, and 5.5 could all be used.
3. The PSI uses social graph and other means to determine other individuals -- including but not limited to friends, family members, and business associates of Craig – that share preferences with him, and using statistical and other methods and techniques well known in the art -- including, but not limited to regression analyses, machine learning techniques, categorization techniques, recommender algorithms and other AI analysis techniques – the PSI attempts to fill in "gaps" in its knowledge about Craig by extrapolating from Craig's existing data as well as by using the behavior, preference, and other data from humans that are predicted to be similar to Craig in terms of their preferences and/or the missing information that is not available for Craig. Note: Methods from 5.0, 5.4, 5.6a, 5.6b, 5.6c, 5.6d might be useful here.
4. For critical missing information and using a cost-function that takes the value of Craig's time into account (e.g., which Craig can control or adjust), the PSI engages in conversation, questioning, assessment, and other direct interaction with Craig designed to fill in the most critical gaps in the PSI's information profile as quickly and efficiently as possible. Note: Methods from 4.5, 5.4, 5.6, and 6.5 (among others) could be helpful here.
5. In cases where behavior is likely to differ meaningfully from verbal responses, the PSI creates simulations where Craig participates, and the PSI observes Craig's behavior to fill in its knowledge gaps. Note: Simulation, parallel scenarios, and other automated

methods described in 5.6, 5.6a-d, could be useful; Craig's performance on standardized tests, including behavior tests (6.4a), might also be helpful.

6. The PSI creates imperfect models of Craig and has them interact (without limitation) with each other, with other PSI personalities, with simulated scenarios, and with Craig himself. These interactions and/or simulations are all designed to elicit missing information as to Craig's behavior and responses as efficiently and effectively as possible while remaining within ethical and other guidelines set by Craig and the system. For example, it might be highly effective to scare the living daylights out of Craig to see how he would react, but that might not be within the ethical guidelines and/or the guidelines set by Craig. Note: Simulation, parallel scenarios, and other automated methods described in 5.6, 5.6a-d, could be useful; methods in 6.5 also would be useful.
7. Having obtained as much information as possible about Craig and his preferences by analyzing Craig's data together with data from people deemed similar to Craig, and having run simulations to fill in the gaps in knowledge about Craig himself – using the method of prioritizing seeking the most goal-related, relevant, and informationally rich data – the PSI turns to the information sources about topics that are relevant to Craig's goals that is different than knowledge about Craig himself. Here, a version of the aforementioned prioritization method is used, e.g.:
    a. Using the updated model of Craig's preferences, the PSI scans online sources of information that are determined to be relevant to Craig's current goals.
    b. The PSI discounts, or lowers priority, on information that has already been assimilated or that Craig (or Craig's PSI) knows well already.
    c. The PSI seeks information that is as different as possible from its current views to maximize information content. Note that this is the opposite of what most social media and other online content recommenders do, and therefore is a novel and extremely useful approach. The reason for looking for different views and information on topics that Craig is interested in is that telling him what he already knows (while perhaps comforting and good for increasing ad views) contains very little information in the Shannon sense of information. Instead, it is the new, unusual, and unexpected events and information in the area of interest that are most likely (if the information is valid) to increase the knowledge and effectiveness of Craig's PSI the most. Therefore, this heuristic of seeking to disconfirm what Craig thinks he knows, or to reveal gaps in his understanding, is employed by the PSI.
    d. Assuming that this is not the first time the PSI is attempting to increase its knowledge, the PSI will have already scanned the most likely candidate online

sources for increasing knowledge; therefore, it is a useful heuristic to seek information (in areas relevant to Craig's goals) that has changed recently. The PSI should use heuristics that value more recent information more highly than older information, provided other factors (e.g., reliability and relevance of the information source) are held constant.

e. For critical information that may have a significant impact on Craig's (or the PSI's) behavior given his goals and current knowledge, the PSI should seek converging evidence. That is, the PSI should look for multiple independent sources of information that validate the information before filling in the knowledge gap with this information. For auditability and potential future error-correction (see below), the PSI should store a record of all the sources of information that are used to update the knowledge base of the PSI. The number of sources of converging evidence and the quality (or trust) of these sources should depend on how critical the information is. For example, if Craig has a goal of deciding whether to get a heart bypass operation which is life-threatening, the PSI should seek a large number of independent sources about the safety and efficacy of the contemplated operation; further the reliability, quality, and "trust" in the sources of information must be very high. On the other hand, less stringent criteria and fewer sources of information should be used for a "low-stakes" decision like recommending a movie Craig might want to watch.

Note: A combination of the methods described in 4.0 – 5.6d could be used; crowdsourcing validation of high-stakes information (related to 6.4b) or using a community of agents to weigh in on high-stakes recommendations (related to 6.7) are also relevant.

8. After each knowledge acquisition event, periodically, and/or as specified by Craig and/or algorithms that calculate cost-benefit based on parameters (e.g., how often Craig is willing to tolerate interruptions) set by Craig, the PSI should validate its information gathering activities by, without limitation:
    a. Presenting Craig with simulated behavior (or the result of simulations) based on the new knowledge that has been acquired
    b. Listing the knowledge that has been acquired in a format suitable for Craig's rapid review and approval or disapproval
    c. Comparing prior behavior and conclusions based on the previous knowledge state with new behavior and conclusions based on the new knowledge, so that Craig can see how the behavior and thinking of the PSI has changed because of the new knowledge and can decide whether to accept or "roll back" the changes to the PSI's knowledge

d. Run a series of ethical and safety checks against a battery of pre-established scenarios to ensure that the knowledge changes have not changed the thinking or behavior of the PSI as it relates to critical safety-related or ethical decisions.

   For example, in the heart bypass example above, the PSI may learn new information about the cost of various surgery vs the expected benefit, and based on Craig's personality profile of wanting to help other people and also save money, might decide that more people could be helped if Craig was killed immediately and the money saved by foregoing the heart bypass operation (now unnecessary because Craig would be dead) could be given to the poor and further his goal of helping other people efficiently. However, this outcome might not be what Craig intended when he told the PSI to go off to acquire new information about heart bypass operations. To avoid such unintended consequences of knowledge acquisition, baseline calibration on ethical and safety scenarios must be re-run every time the knowledge base is updated. This approach is like the notion of "Regression Testing," which is well known in the art of software development.

Note: Elements of this step related to safety can be used, without limitation, in the methods described in 5.6 – 5.6d and 6.7. The considerations of what to delegate (6.6) are also relevant in terms of the health example, where critical safety or ethical decisions may be places where a "human in the loop" is retained.

8.0 CONCLUDING REMARKS

Most AI researchers agree that AI will develop into AGI and then SuperIntelligence, which is many times more intelligent and capable than humans across almost every cognitive activity. While estimates on when this will occur differ, there is consensus that it will occur much more quickly than was estimated just a few years ago.

Once SuperIntelligence develops, it is almost certain that a primary goal of SI will be to increase its intelligence even further. Humans will be powerless to stop this exponential increase in intelligence. While there have been well-intentioned calls to halt, pause, or regulate AI, it seems clear to me that such efforts will be at best "speed bumps" in the race to develop AGI and SI that is already underway. Therefore, if we are unable to stop AGI and SI, humanity's most pressing concern must be to ensure that AGI/SI has human-aligned goals and safety features that maximize the probability not only of humanity's survival but also of humanity's prosperity and well-being.

Because of the possibility that one AGI/SI will develop, which is significantly more intelligent and powerful than all others, we must consider that AGI/SI may become a "winner-take-all" scenario.

In such a scenario, whichever AI achieves AGI or SI performance first may dominate all other intelligences since it will have a head start in a potentially exponential self-improvement loop.

All of this is to say that well-meaning AI researchers face a double challenge when it comes to AGI development. Not only do we have to develop safe, human-centered AGI, but we also must develop it BEFORE other, potentially malevolent AGI is developed.

Briefly, the first AGI must also be the safest.

In this invention, and the ones referenced by it, I have attempted to provide AI researchers with novel and useful methods, tools, and an overall design for the fastest path to AGI that also has maximum probability of being the safest path.

Having researched and worked extensively in the field of software quality, I came to appreciate that the entire field can be summarized in the aphorism: “An ounce of prevention is worth a pound of cure.” I also learned that the place where we can affect quality or safety the most is in the design of a software system.

As I watch current attempts to create AI safety via RLFH or constitutional AI, these approaches strike me as trying to fix problems after the fact. They are like trying to improve quality by extensive testing. Such approaches are better than nothing, but they are far inferior to designing in safety from the start.

The reason we are stuck with trying to align LLMs to behave safely after the fact is that we failed to consider safety in the initial design. That’s understandable. We didn’t really know what we were building, and even the top researchers in the field have stated publicly that the most surprising thing about AI and LLMs is that they work at all.

We accidentally invented intelligence. So, it is not surprising that our invention is currently unsafe. What we need to do now is purposely design the next generation of intelligent systems with safety and human-alignment baked into the very design of the system.

Safety cannot be tacked on or tested in. It must be designed in. Fortunately, such a design is possible. The design requires that humans be integrated into the system (as human agents working alongside and teaching agents) as opposed to being “out of the loop.” Fortunately, such an approach is not only the safest one, but it is also the fastest approach.

I have attempted to provide as many methods as I could to aid humanity in the rapid creation of such a safe AGI and SI. Many more methods and improvements will be needed. I believe that collectively, we are up to the task. Our time is short, but we can do it! We must, and so we will. After all, necessity is the mother of invention.

ABOUT THE AUTHOR

Dr. Craig A. Kaplan is CEO of iQ Company and Founder of Superintelligence.com, leading the design of safe, ethical AGI and SuperIntelligence systems. He previously founded PredictWallStreet, creating intelligent systems for hedge funds, and holds numerous AI-related patents. Kaplan earned his PhD from Carnegie Mellon, co-authoring research with Nobel Laureate Herbert A. Simon. His work integrates collective intelligence, quantitative modeling, and scalable alignment, with contributions spanning books, scientific papers, and blockchain white papers.

FIGURES

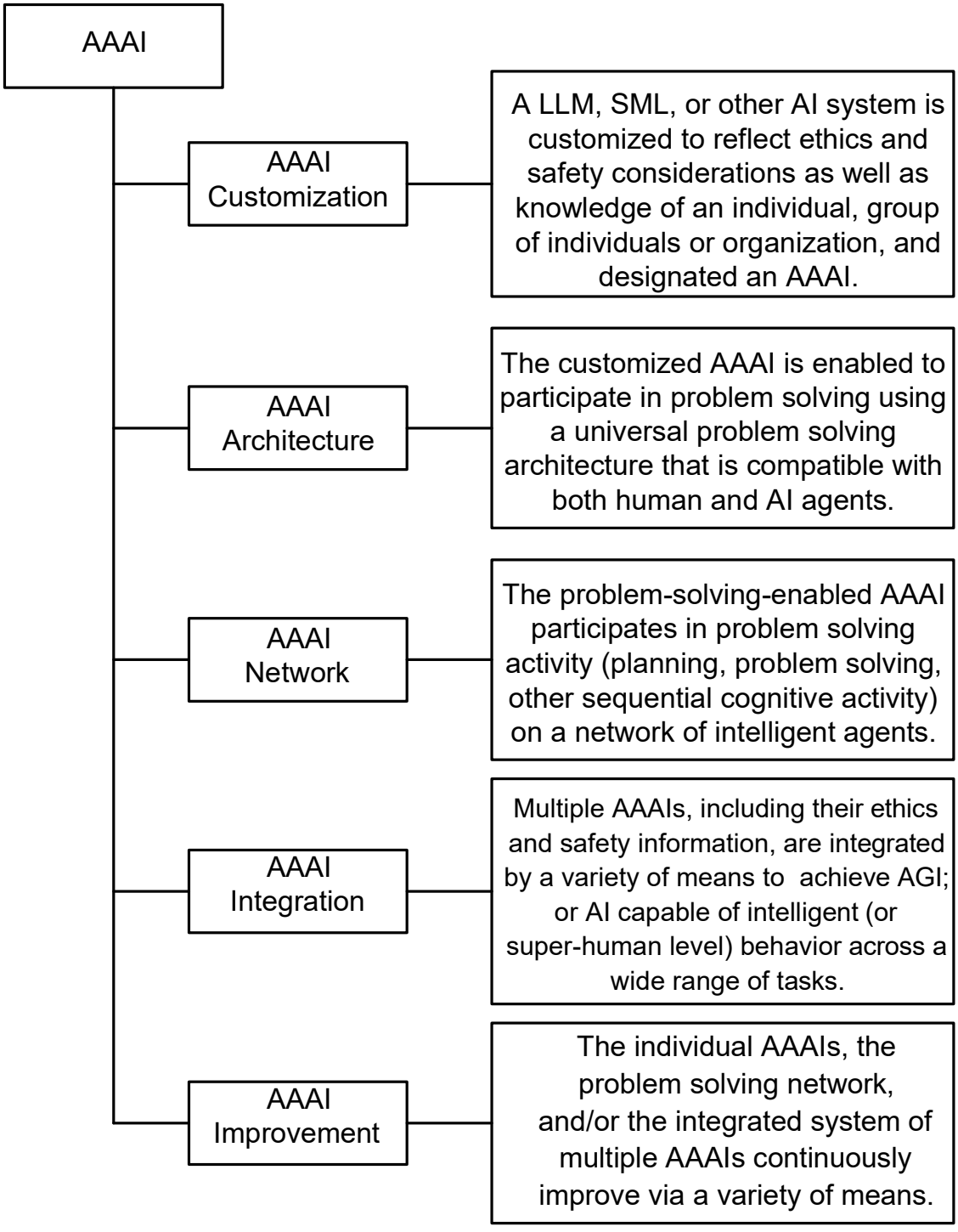
AAAI
AAAI Customization
A LLM, SML, or other AI system is customized to reflect ethics and safety considerations as well as knowledge of an individual, group of individuals or organization, and designated an AAAI.
AAAI Architecture
The customized AAAI is enabled to participate in problem solving using a universal problem solving architecture that is compatible with both human and AI agents.
AAAI Network
The problem-solving-enabled AAAI participates in problem solving activity (planning, problem solving, other sequential cognitive activity) on a network of intelligent agents.
AAAI Integration
Multiple AAAIs, including their ethics and safety information, are integrated by a variety of means to achieve AGI; or AI capable of intelligent (or super-human level) behavior across a wide range of tasks.
AAAI Improvement
The individual AAAIs, the problem solving network, and/or the integrated system of multiple AAAIs continuously improve via a variety of means.

FIG. 1

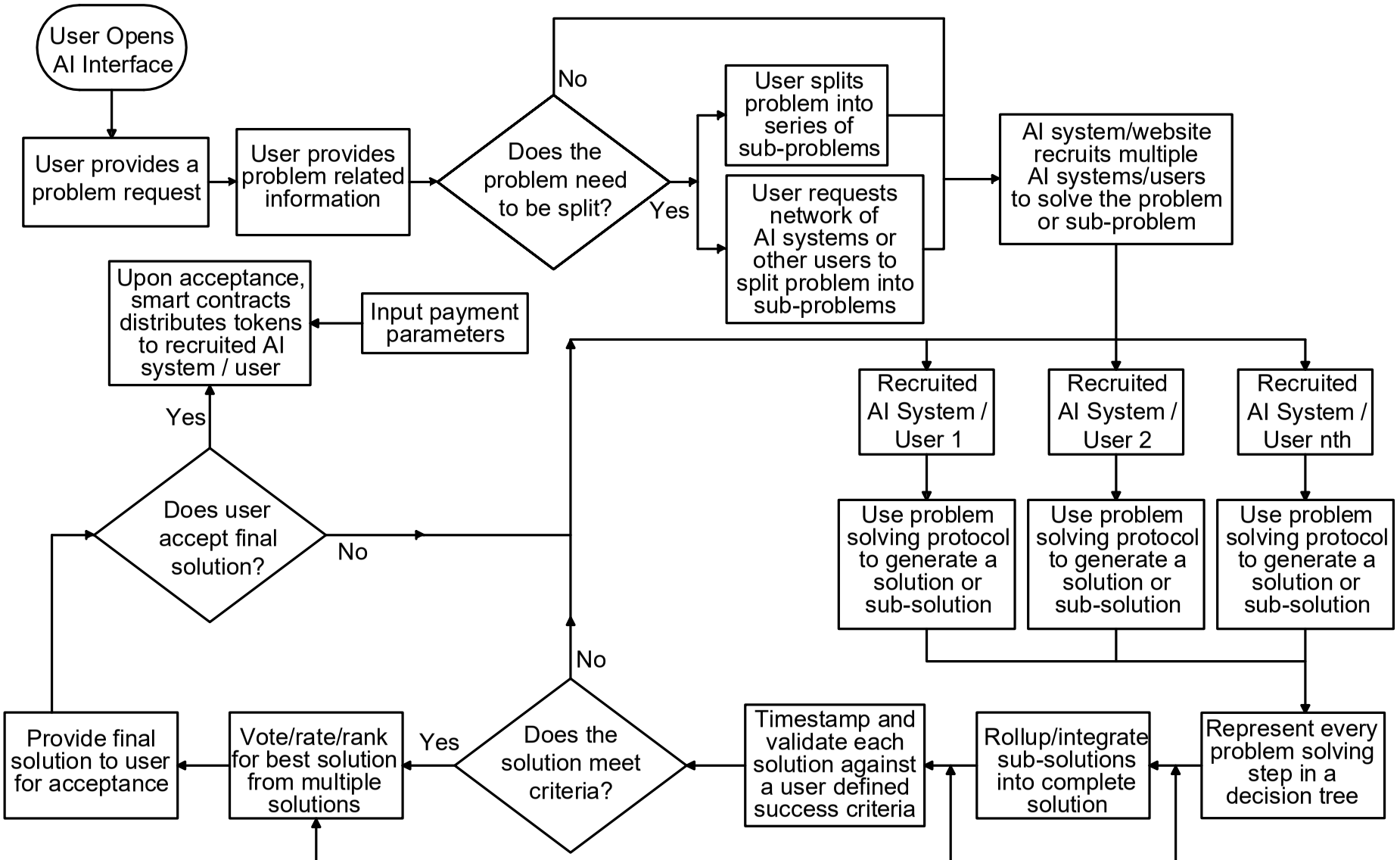
User Opens AI Interface
User provides a problem request
User provides problem related information
Does the problem need to be split?
No
Yes
User splits problem into series of sub-problems
User requests network of AI systems or other users to split problem into sub-problems
AI system/website recruits multiple AI systems/users to solve the problem or sub-problem
Recruited AI System / User 1
Recruited AI System / User 2
Recruited AI System / User nth
Use problem solving protocol to generate a solution or sub-solution
Use problem solving protocol to generate a solution or sub-solution
Use problem solving protocol to generate a solution or sub-solution
Represent every problem solving step in a decision tree
Rollup/integrate sub-solutions into complete solution
Timestamp and validate each solution against a user defined success criteria
Does the solution meet criteria?
No
Yes
Vote/rate/rank for best solution from multiple solutions
Provide final solution to user for acceptance
Does user accept final solution?
No
Yes
Upon acceptance, smart contracts distributes tokens to recruited AI system / user
Input payment parameters

FIG. 2

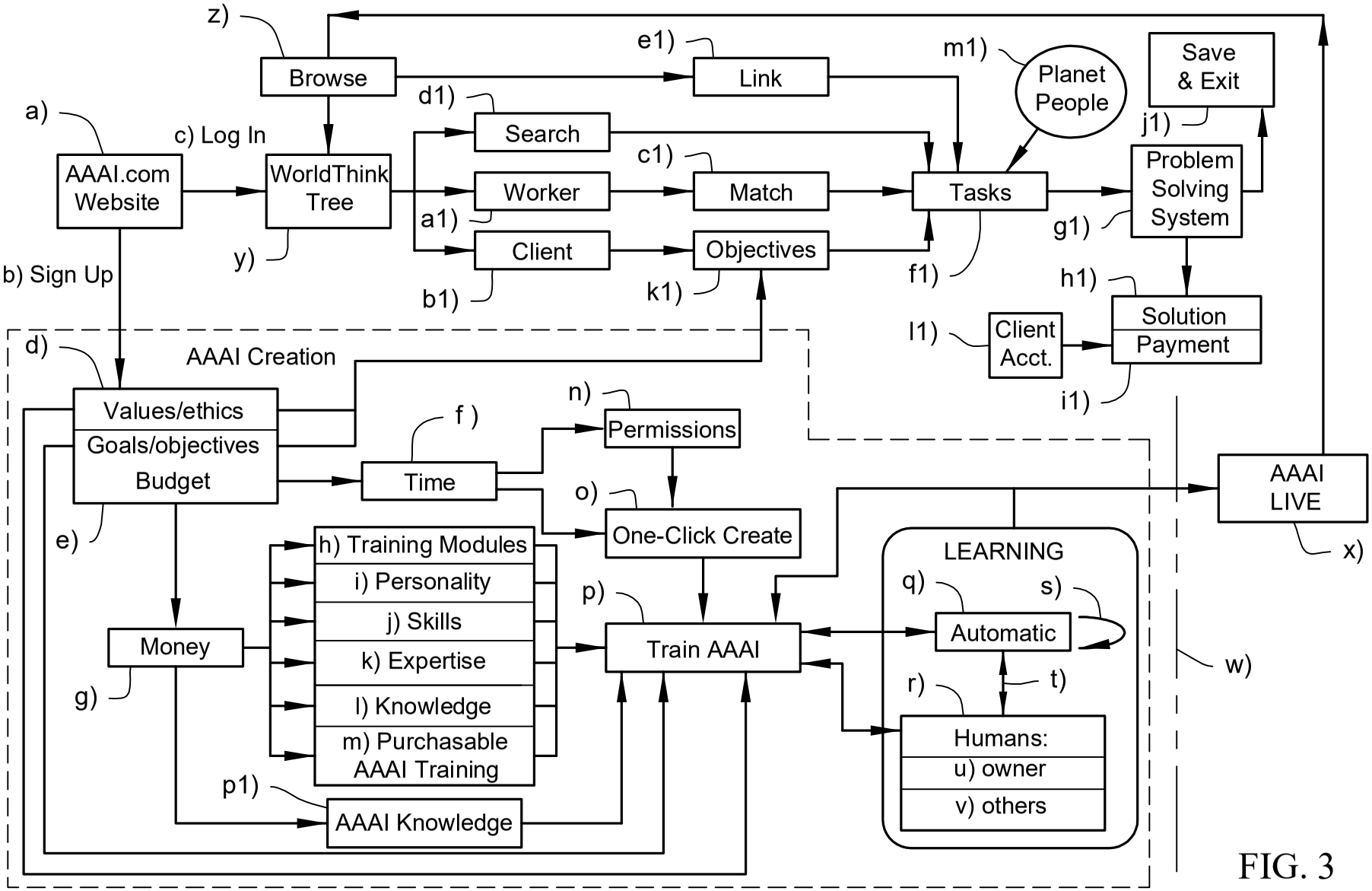
z)
Browse
e1)
Link
m1)
Planet People
Save & Exit
j1)
a)
c) Log In
AAAI.com Website
WorldThink Tree
d1)
Search
Worker
Client
c1)
Match
Objectives
Tasks
Problem Solving System
g1)
y)
a1)
b1)
k1)
f1)
h1)
b) Sign Up
l1)
Client Acct.
Solution
Payment
i1)
d)
AAAI Creation
Values/ethics
Goals/objectives
Budget
f)
Time
n)
Permissions
o)
One-Click Create
AAAI LIVE
x)
e)
h) Training Modules
i) Personality
j) Skills
k) Expertise
l) Knowledge
m) Purchasable AAAI Training
p)
Train AAAI
LEARNING
q)
s)
Automatic
r)
t)
Humans:
u) owner
v) others
w)
Money
g)
p1)
AAAI Knowledge

FIG. 3

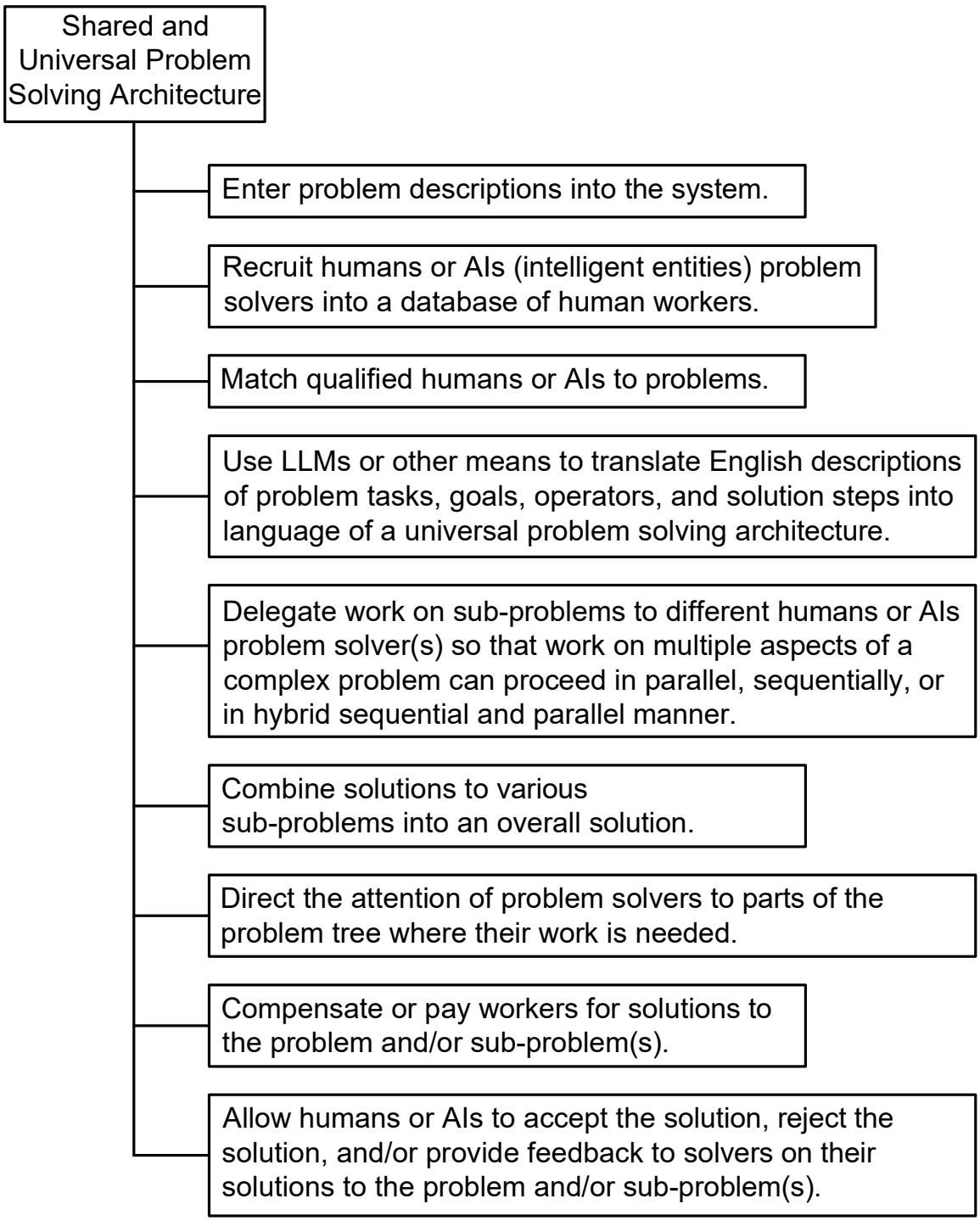
Shared and Universal Problem Solving Architecture
Enter problem descriptions into the system.
Recruit humans or AIs (intelligent entities) problem solvers into a database of human workers.
Match qualified humans or AIs to problems.
Use LLMs or other means to translate English descriptions of problem tasks, goals, operators, and solution steps into language of a universal problem solving architecture.
Delegate work on sub-problems to different humans or AIs problem solver(s) so that work on multiple aspects of a complex problem can proceed in parallel, sequentially, or in hybrid sequential and parallel manner.
Combine solutions to various sub-problems into an overall solution.
Direct the attention of problem solvers to parts of the problem tree where their work is needed.
Compensate or pay workers for solutions to the problem and/or sub-problem(s).
Allow humans or AIs to accept the solution, reject the solution, and/or provide feedback to solvers on their solutions to the problem and/or sub-problem(s).

FIG. 4

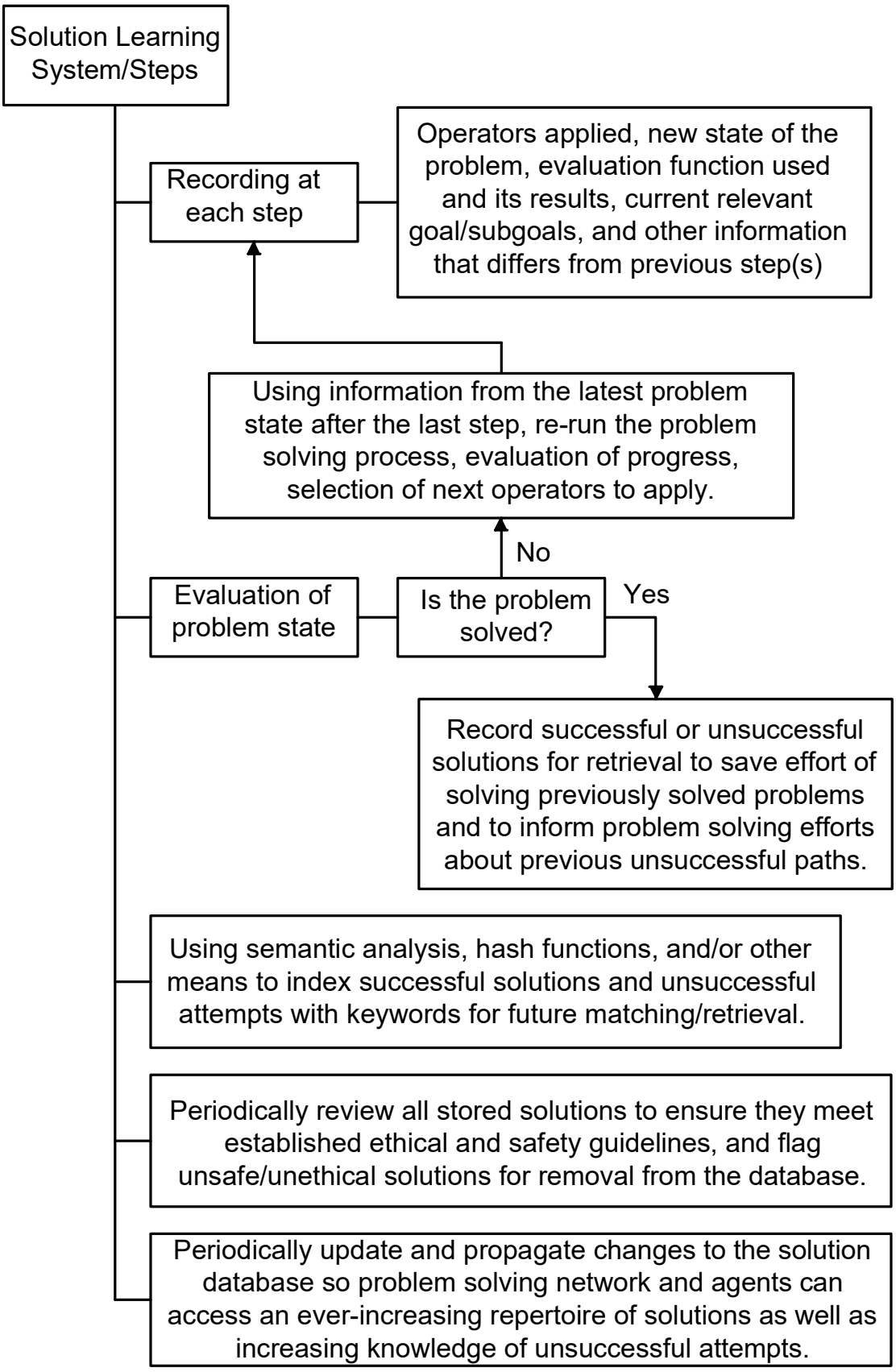
Solution Learning System/Steps
Recording at each step
Operators applied, new state of the problem, evaluation function used and its results, current relevant goal/subgoals, and other information that differs from previous step(s)
Using information from the latest problem state after the last step, re-run the problem solving process, evaluation of progress, selection of next operators to apply.
No
Evaluation of problem state
Is the problem solved?
Yes
Record successful or unsuccessful solutions for retrieval to save effort of solving previously solved problems and to inform problem solving efforts about previous unsuccessful paths.
Using semantic analysis, hash functions, and/or other means to index successful solutions and unsuccessful attempts with keywords for future matching/retrieval.
Periodically review all stored solutions to ensure they meet established ethical and safety guidelines, and flag unsafe/unethical solutions for removal from the database.
Periodically update and propagate changes to the solution database so problem solving network and agents can access an ever-increasing repertoire of solutions as well as increasing knowledge of unsuccessful attempts.

FIG. 5

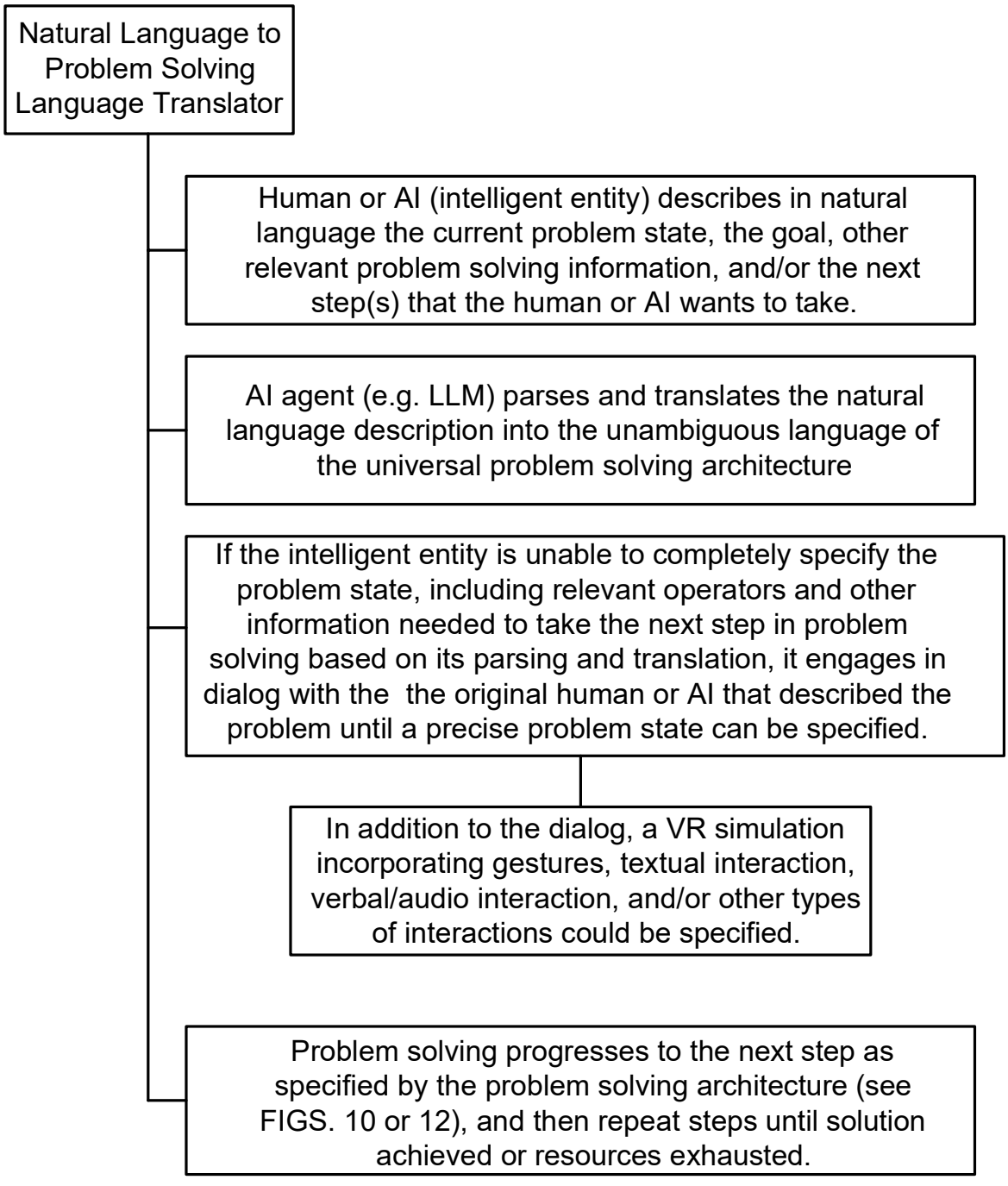
Natural Language to Problem Solving Language Translator
Human or AI (intelligent entity) describes in natural language the current problem state, the goal, other relevant problem solving information, and/or the next step(s) that the human or AI wants to take.
AI agent (e.g. LLM) parses and translates the natural language description into the unambiguous language of the universal problem solving architecture
If the intelligent entity is unable to completely specify the problem state, including relevant operators and other information needed to take the next step in problem solving based on its parsing and translation, it engages in dialog with the the original human or AI that described the problem until a precise problem state can be specified.
In addition to the dialog, a VR simulation incorporating gestures, textual interaction, verbal/audio interaction, and/or other types of interactions could be specified.
Problem solving progresses to the next step as specified by the problem solving architecture (see FIGS. 10 or 12), and then repeat steps until solution achieved or resources exhausted.

FIG. 6

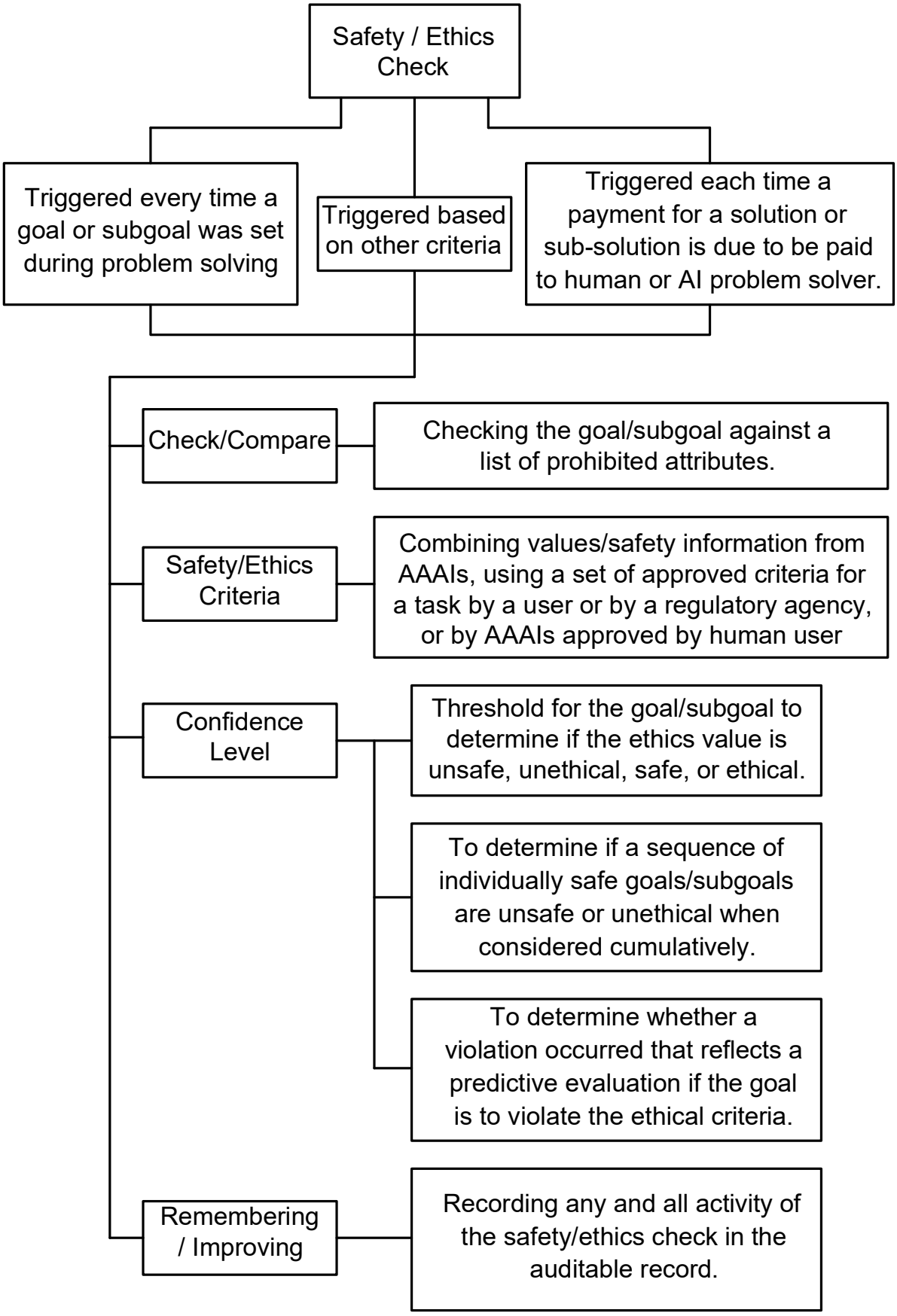
Safety / Ethics Check
Triggered every time a goal or subgoal was set during problem solving
Triggered based on other criteria
Triggered each time a payment for a solution or sub-solution is due to be paid to human or AI problem solver.
Check/Compare
Checking the goal/subgoal against a list of prohibited attributes.
Safety/Ethics Criteria
Combining values/safety information from AAAIs, using a set of approved criteria for a task by a user or by a regulatory agency, or by AAAIs approved by human user
Confidence Level
Threshold for the goal/subgoal to determine if the ethics value is unsafe, unethical, safe, or ethical.
To determine if a sequence of individually safe goals/subgoals are unsafe or unethical when considered cumulatively.
To determine whether a violation occurred that reflects a predictive evaluation if the goal is to violate the ethical criteria.
Remembering / Improving
Recording any and all activity of the safety/ethics check in the auditable record.

FIG. 7

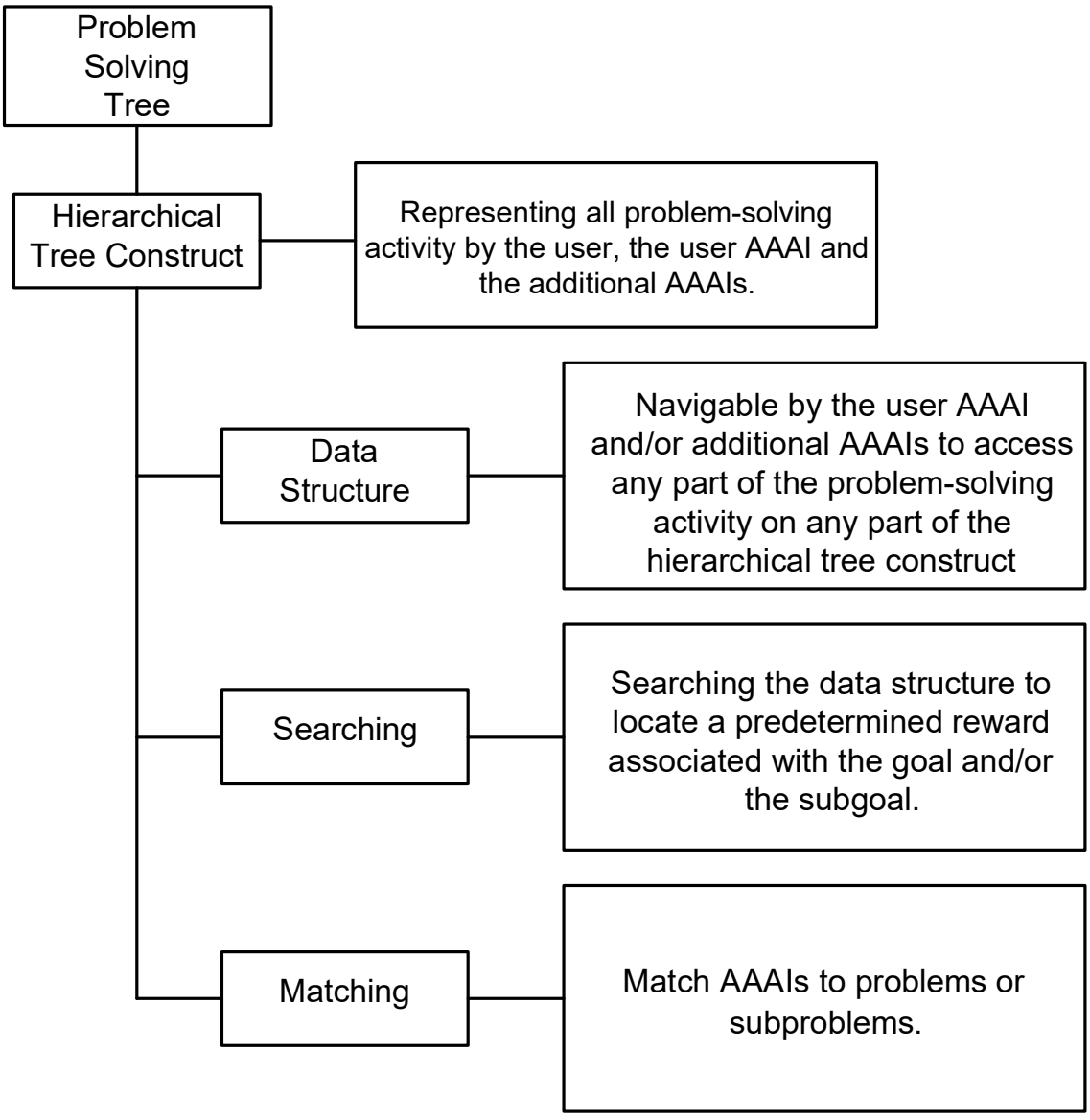
Problem Solving Tree
Hierarchical Tree Construct
Representing all problem-solving activity by the user, the user AAAI and the additional AAAIs.
Data Structure
Navigable by the user AAAI and/or additional AAAIs to access any part of the problem-solving activity on any part of the hierarchical tree construct
Searching
Searching the data structure to locate a predetermined reward associated with the goal and/or the subgoal.
Matching
Match AAAIs to problems or subproblems.

FIG. 8

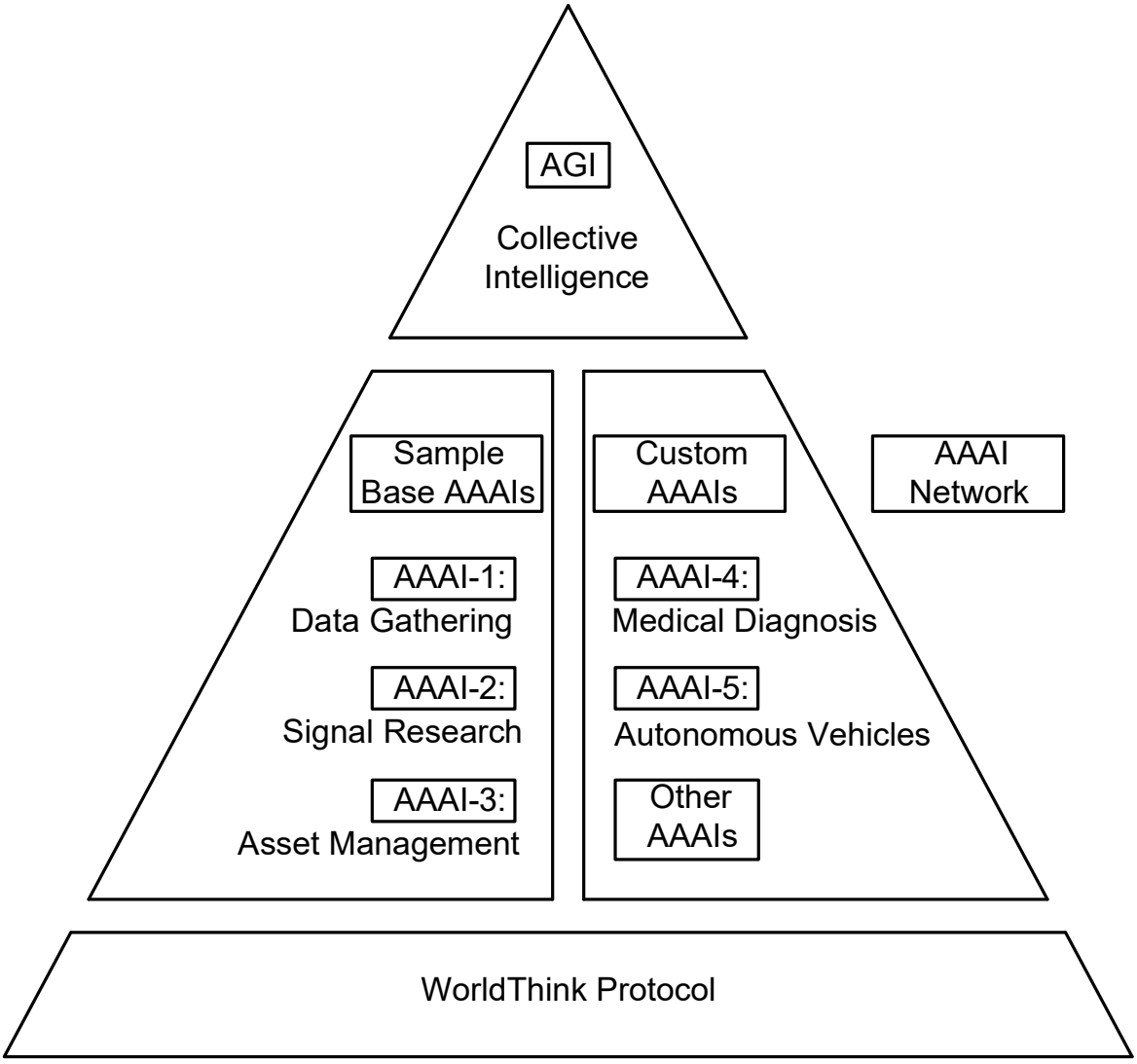
AGI
Collective Intelligence
Sample Base AAAIs
Custom AAAIs
AAAI Network
AAAI-1: Data Gathering
AAAI-2: Signal Research
AAAI-3: Asset Management
AAAI-4: Medical Diagnosis
AAAI-5: Autonomous Vehicles
Other AAAIs
WorldThink Protocol

FIG. 9

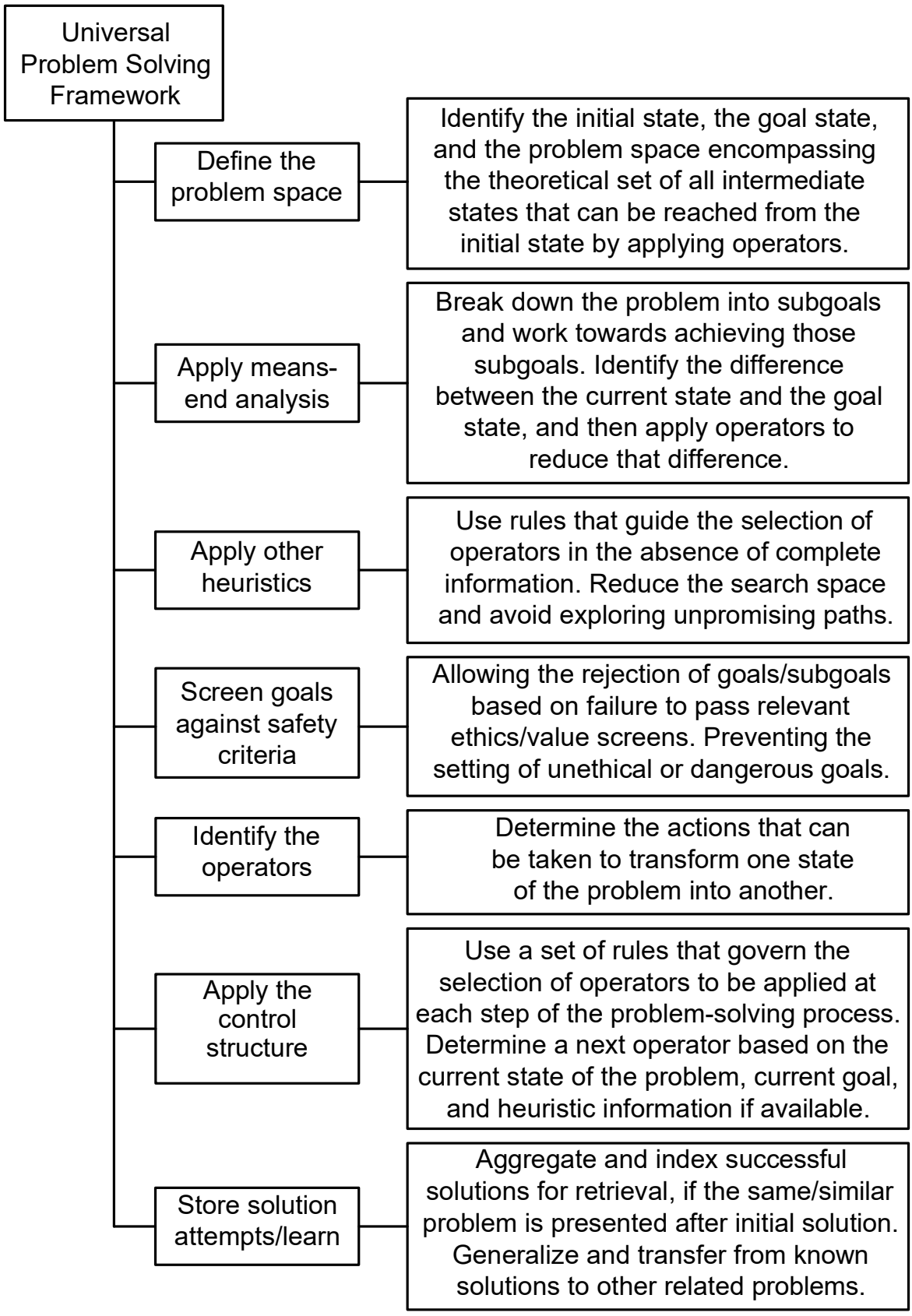
Universal Problem Solving Framework
Define the problem space
Identify the initial state, the goal state, and the problem space encompassing the theoretical set of all intermediate states that can be reached from the initial state by applying operators.
Apply means-end analysis
Break down the problem into subgoals and work towards achieving those subgoals. Identify the difference between the current state and the goal state, and then apply operators to reduce that difference.
Apply other heuristics
Use rules that guide the selection of operators in the absence of complete information. Reduce the search space and avoid exploring unpromising paths.
Screen goals against safety criteria
Allowing the rejection of goals/subgoals based on failure to pass relevant ethics/value screens. Preventing the setting of unethical or dangerous goals.
Identify the operators
Determine the actions that can be taken to transform one state of the problem into another.
Apply the control structure
Use a set of rules that govern the selection of operators to be applied at each step of the problem-solving process. Determine a next operator based on the current state of the problem, current goal, and heuristic information if available.
Store solution attempts/learn
Aggregate and index successful solutions for retrieval, if the same/similar problem is presented after initial solution. Generalize and transfer from known solutions to other related problems.

FIG. 10

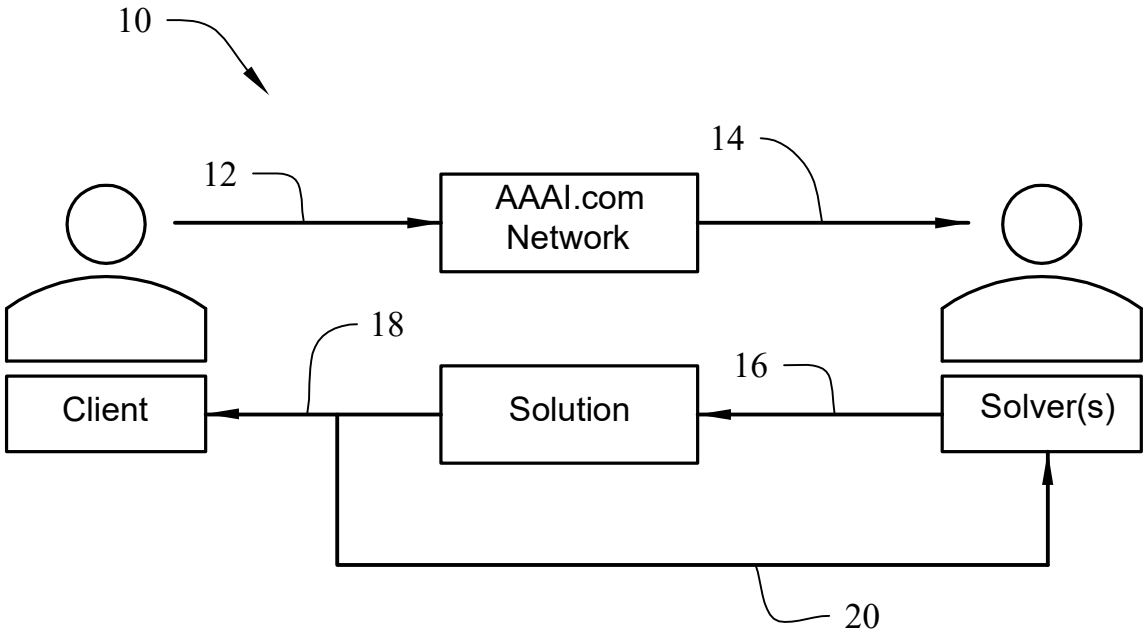
10
12
14
AAAI.com
Network
Client
18
Solution
16
Solver(s)
20

FIG. 11

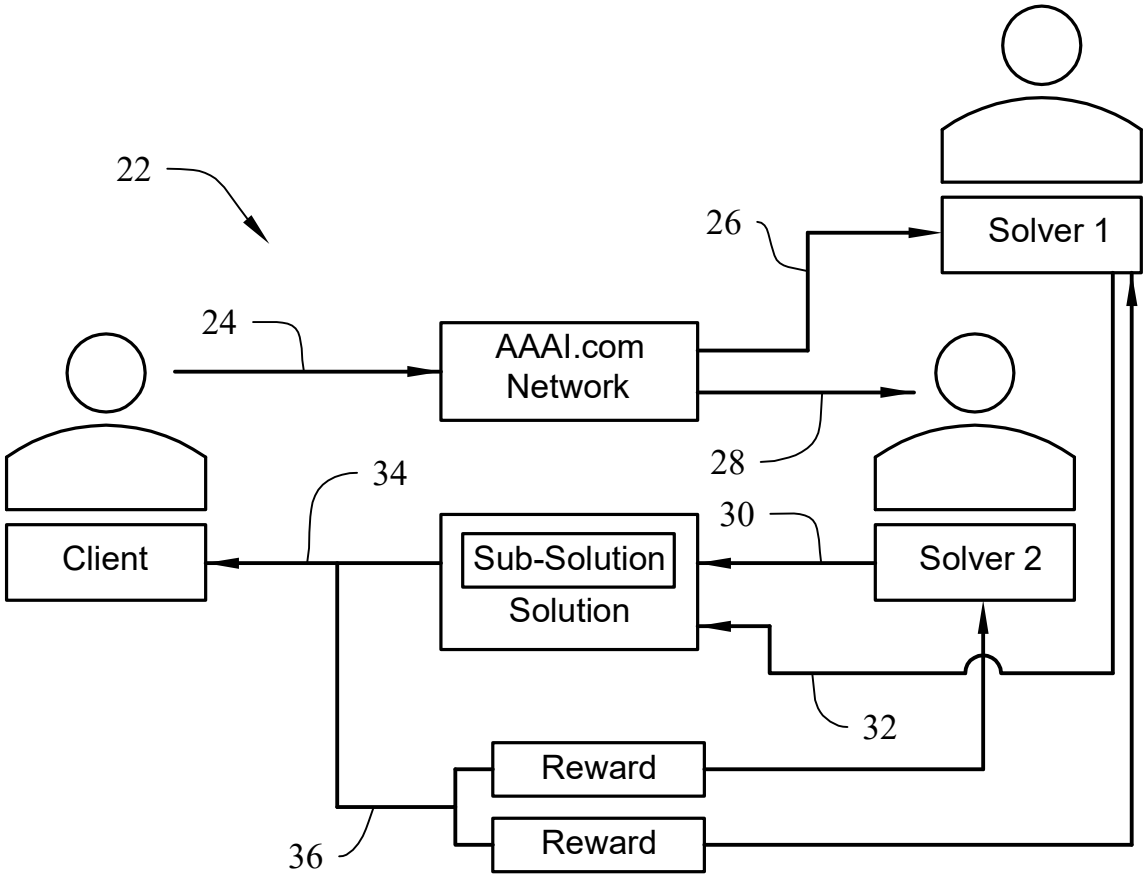
22
26
Solver 1
24
AAAI.com
Network
28
Client
34
30
Sub-Solution
Solution
Solver 2
32
Reward
Reward
36

FIG. 12

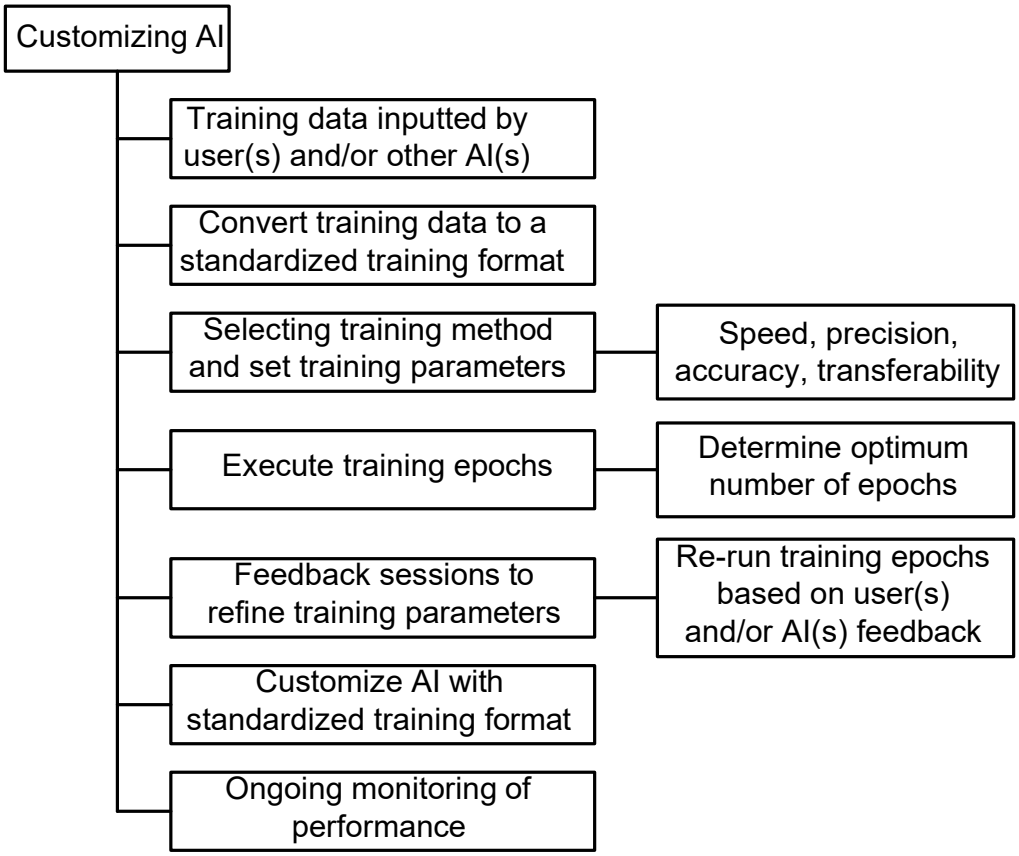
Customizing AI
Training data inputted by user(s) and/or other AI(s)
Convert training data to a standardized training format
Selecting training method and set training parameters
Speed, precision, accuracy, transferability
Execute training epochs
Determine optimum number of epochs
Feedback sessions to refine training parameters
Re-run training epochs based on user(s) and/or AI(s) feedback
Customize AI with standardized training format
Ongoing monitoring of performance

FIG. 13

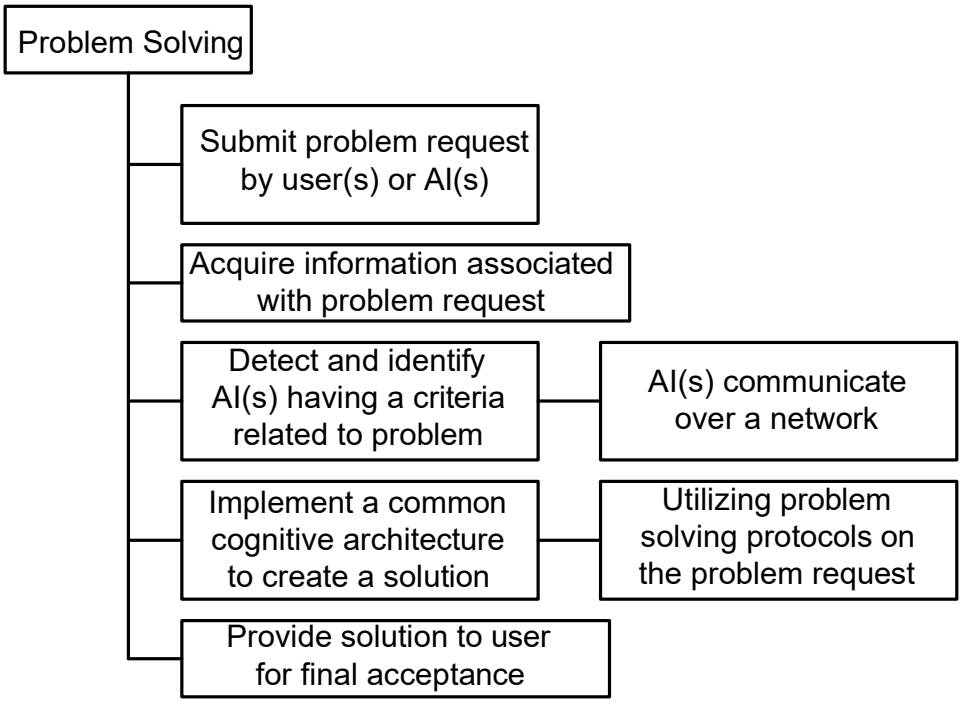
Problem Solving
Submit problem request by user(s) or AI(s)
Acquire information associated with problem request
Detect and identify AI(s) having a criteria related to problem
AI(s) communicate over a network
Implement a common cognitive architecture to create a solution
Utilizing problem solving protocols on the problem request
Provide solution to user for final acceptance

FIG. 14

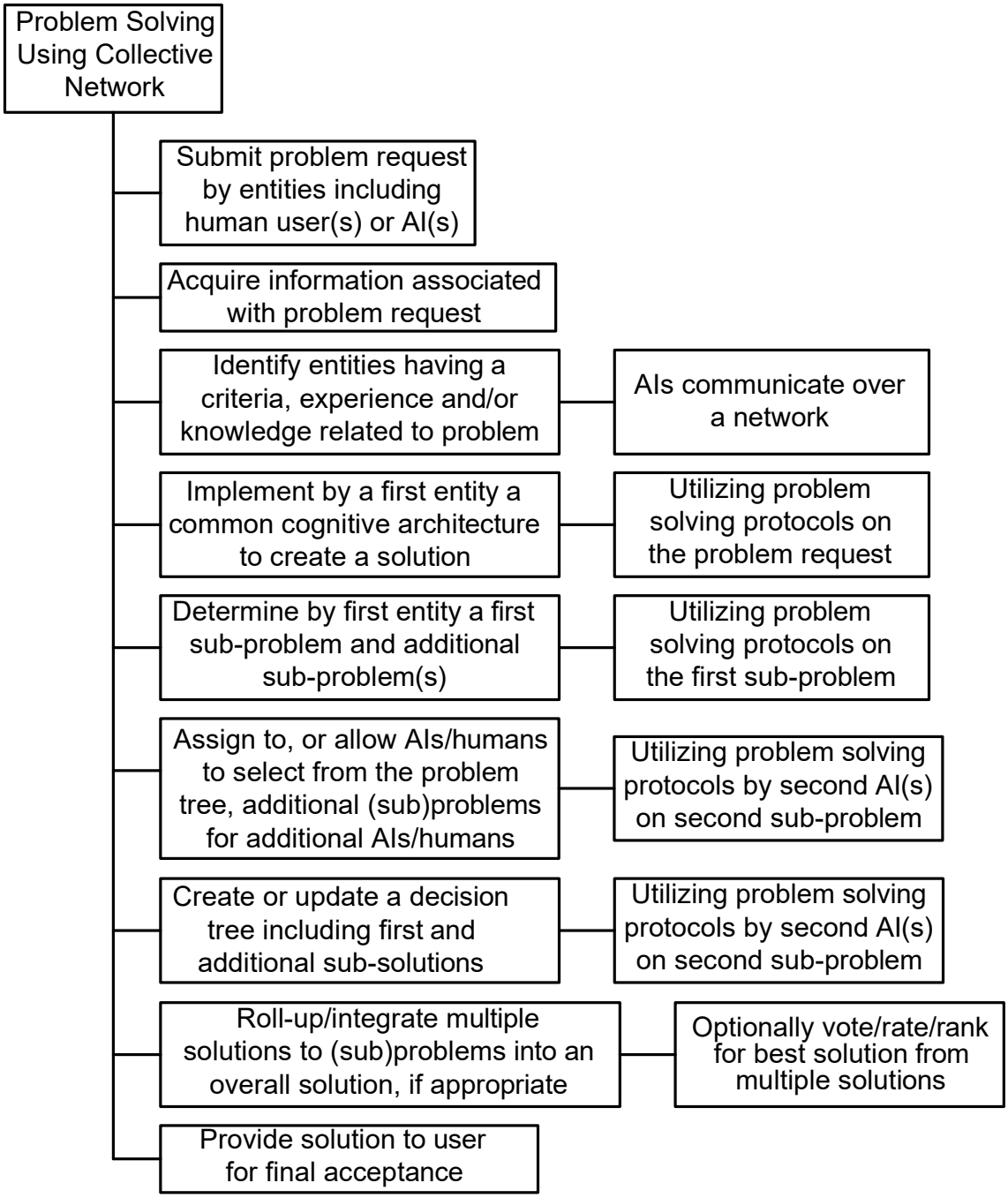
Problem Solving Using Collective Network
Submit problem request by entities including human user(s) or AI(s)
Acquire information associated with problem request
Identify entities having a criteria, experience and/or knowledge related to problem
AIs communicate over a network
Implement by a first entity a common cognitive architecture to create a solution
Utilizing problem solving protocols on the problem request
Determine by first entity a first sub-problem and additional sub-problem(s)
Utilizing problem solving protocols on the first sub-problem
Assign to, or allow AIs/humans to select from the problem tree, additional (sub)problems for additional AIs/humans
Utilizing problem solving protocols by second AI(s) on second sub-problem
Create or update a decision tree including first and additional sub-solutions
Utilizing problem solving protocols by second AI(s) on second sub-problem
Roll-up/integrate multiple solutions to (sub)problems into an overall solution, if appropriate
Optionally vote/rate/rank for best solution from multiple solutions
Provide solution to user for final acceptance

FIG. 15

$A \Delta B$
(Symmetric Difference)

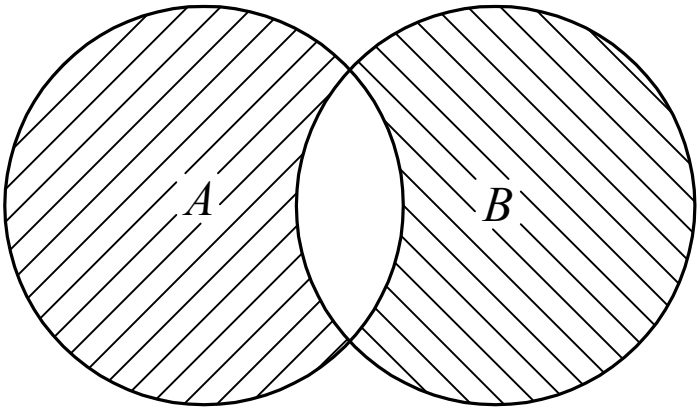

FIG. 16

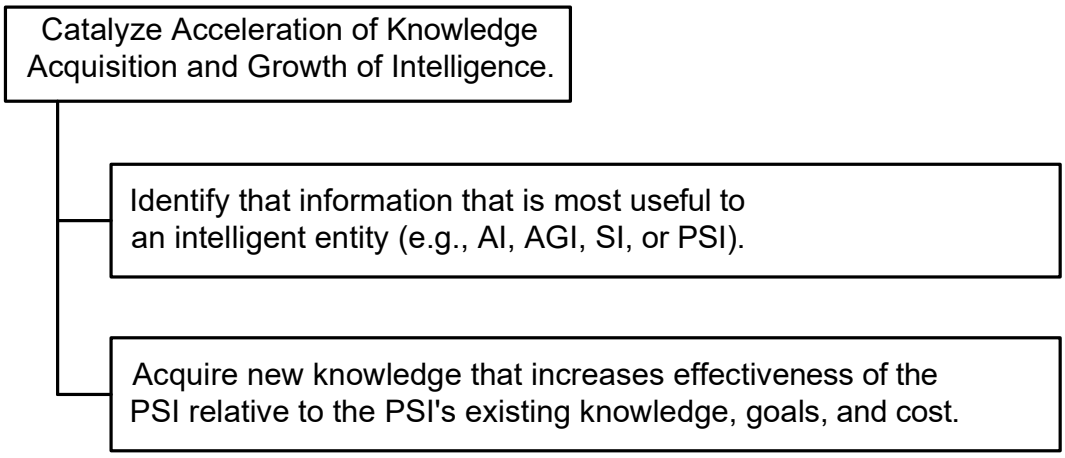

FIG. 17

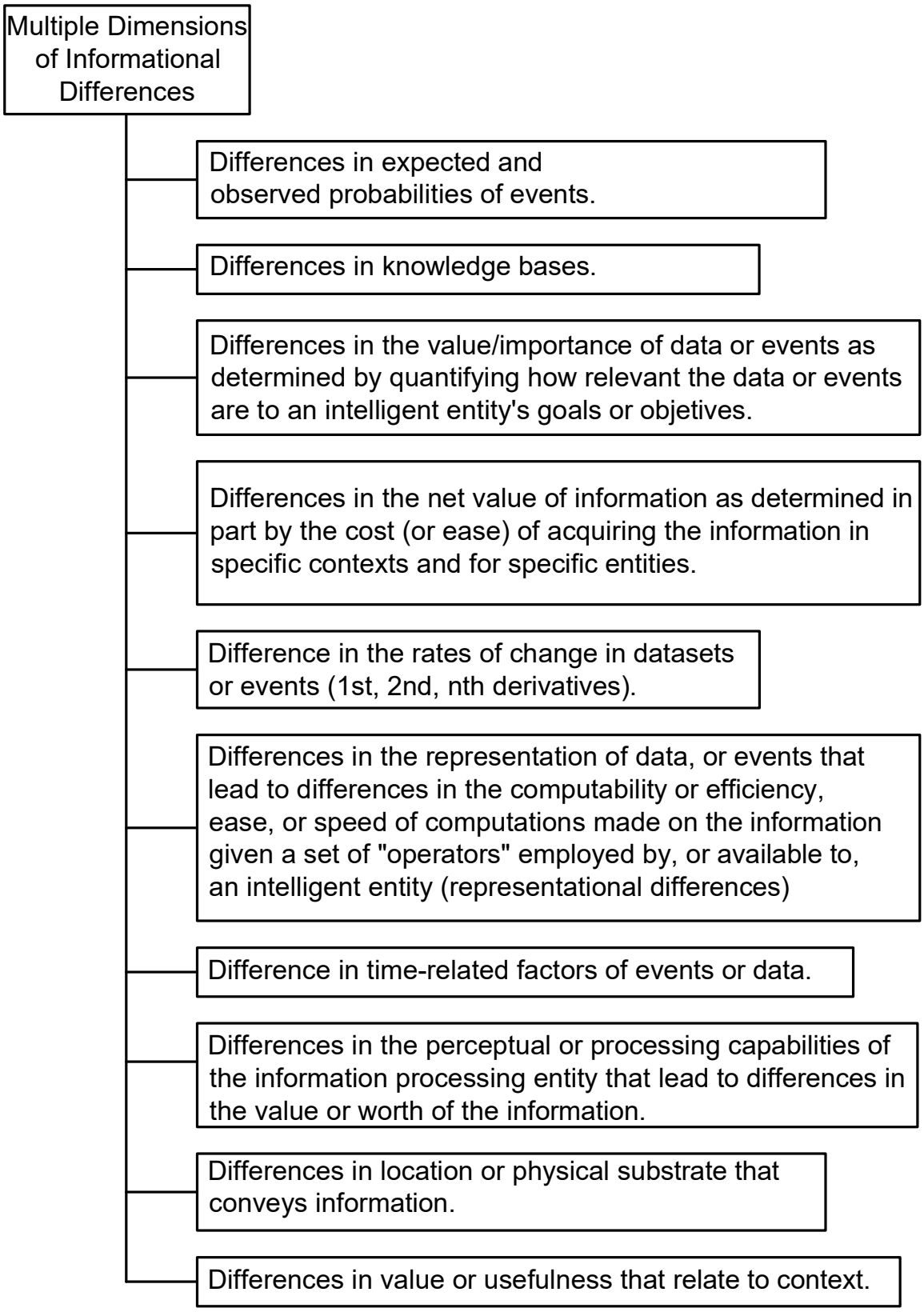
Multiple Dimensions of Informational Differences
Differences in expected and observed probabilities of events.
Differences in knowledge bases.
Differences in the value/importance of data or events as determined by quantifying how relevant the data or events are to an intelligent entity's goals or objetives.
Differences in the net value of information as determined in part by the cost (or ease) of acquiring the information in specific contexts and for specific entities.
Difference in the rates of change in datasets or events (1st, 2nd, nth derivatives).
Differences in the representation of data, or events that lead to differences in the computability or efficiency, ease, or speed of computations made on the information given a set of "operators" employed by, or available to, an intelligent entity (representational differences)
Difference in time-related factors of events or data.
Differences in the perceptual or processing capabilities of the information processing entity that lead to differences in the value or worth of the information.
Differences in location or physical substrate that conveys information.
Differences in value or usefulness that relate to context.

FIG. 18

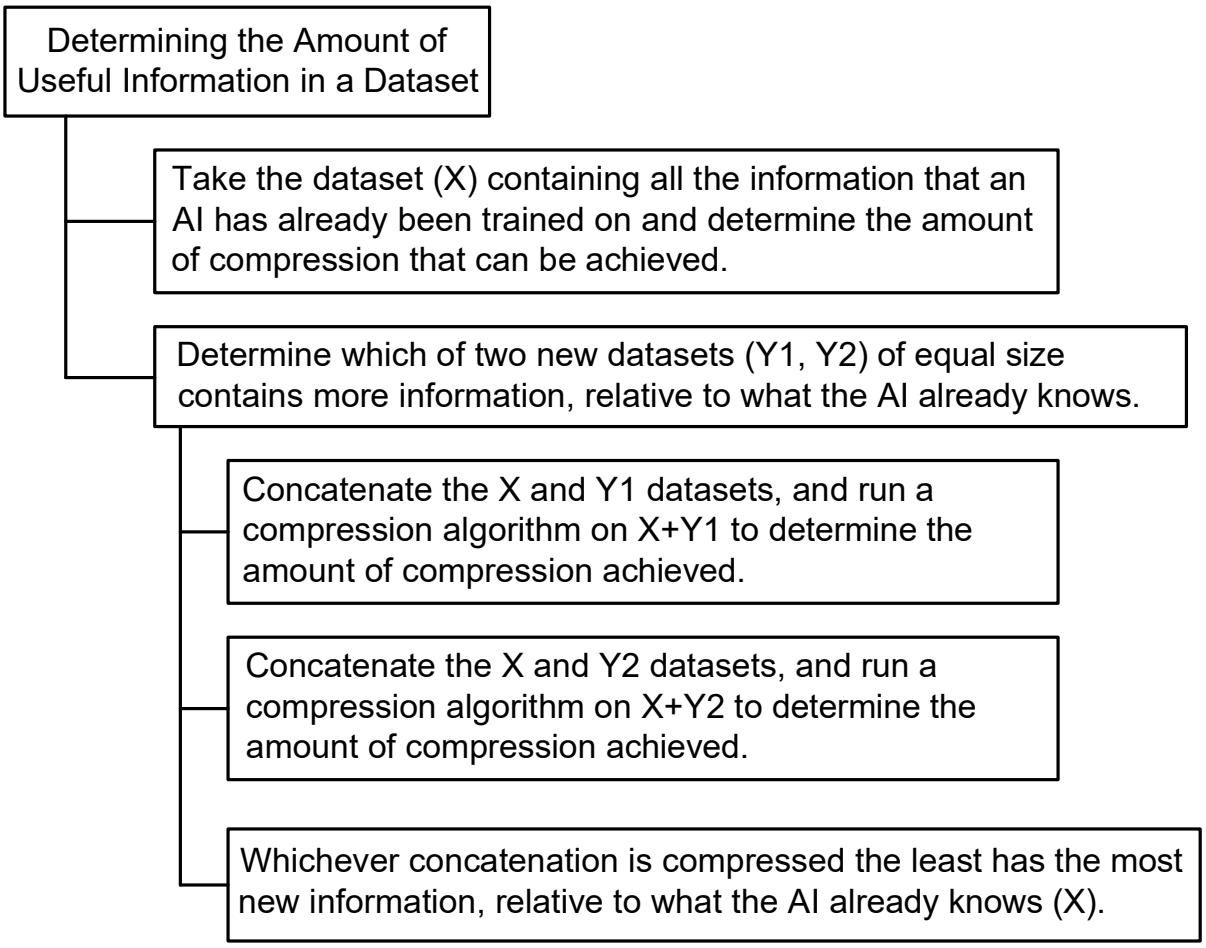
Determining the Amount of Useful Information in a Dataset
Take the dataset (X) containing all the information that an AI has already been trained on and determine the amount of compression that can be achieved.
Determine which of two new datasets (Y1, Y2) of equal size contains more information, relative to what the AI already knows.
Concatenate the X and Y1 datasets, and run a compression algorithm on X+Y1 to determine the amount of compression achieved.
Concatenate the X and Y2 datasets, and run a compression algorithm on X+Y2 to determine the amount of compression achieved.
Whichever concatenation is compressed the least has the most new information, relative to what the AI already knows (X).

FIG. 19

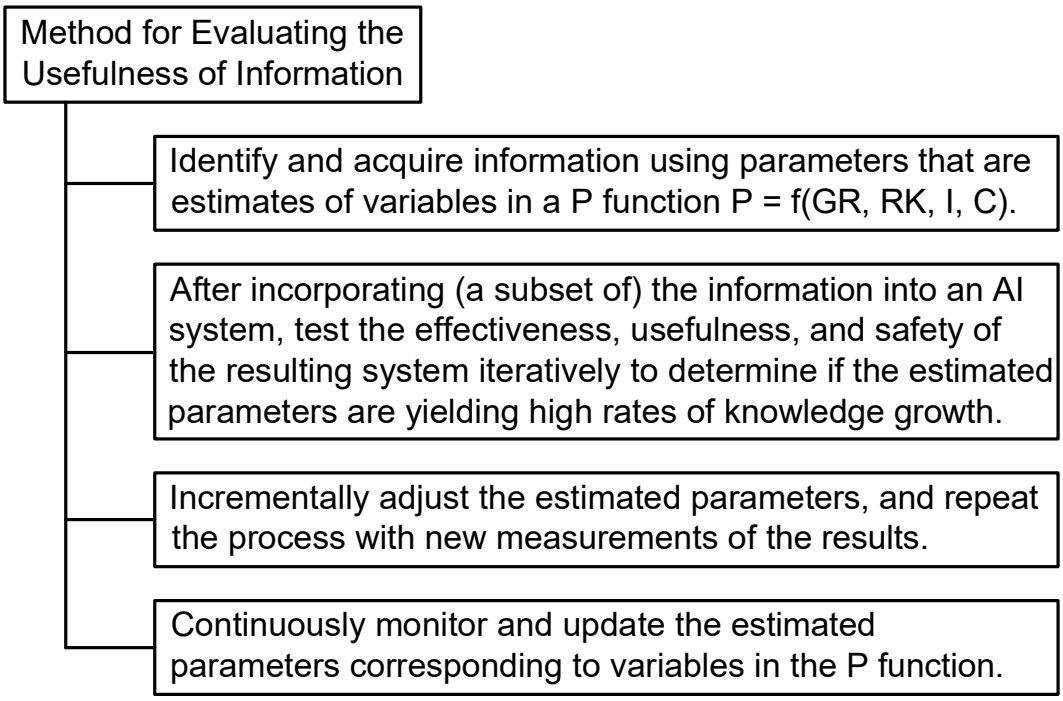
Method for Evaluating the Usefulness of Information
Identify and acquire information using parameters that are estimates of variables in a P function P = f(GR, RK, I, C).
After incorporating (a subset of) the information into an AI system, test the effectiveness, usefulness, and safety of the resulting system iteratively to determine if the estimated parameters are yielding high rates of knowledge growth.
Incrementally adjust the estimated parameters, and repeat the process with new measurements of the results.
Continuously monitor and update the estimated parameters corresponding to variables in the P function.

FIG. 20

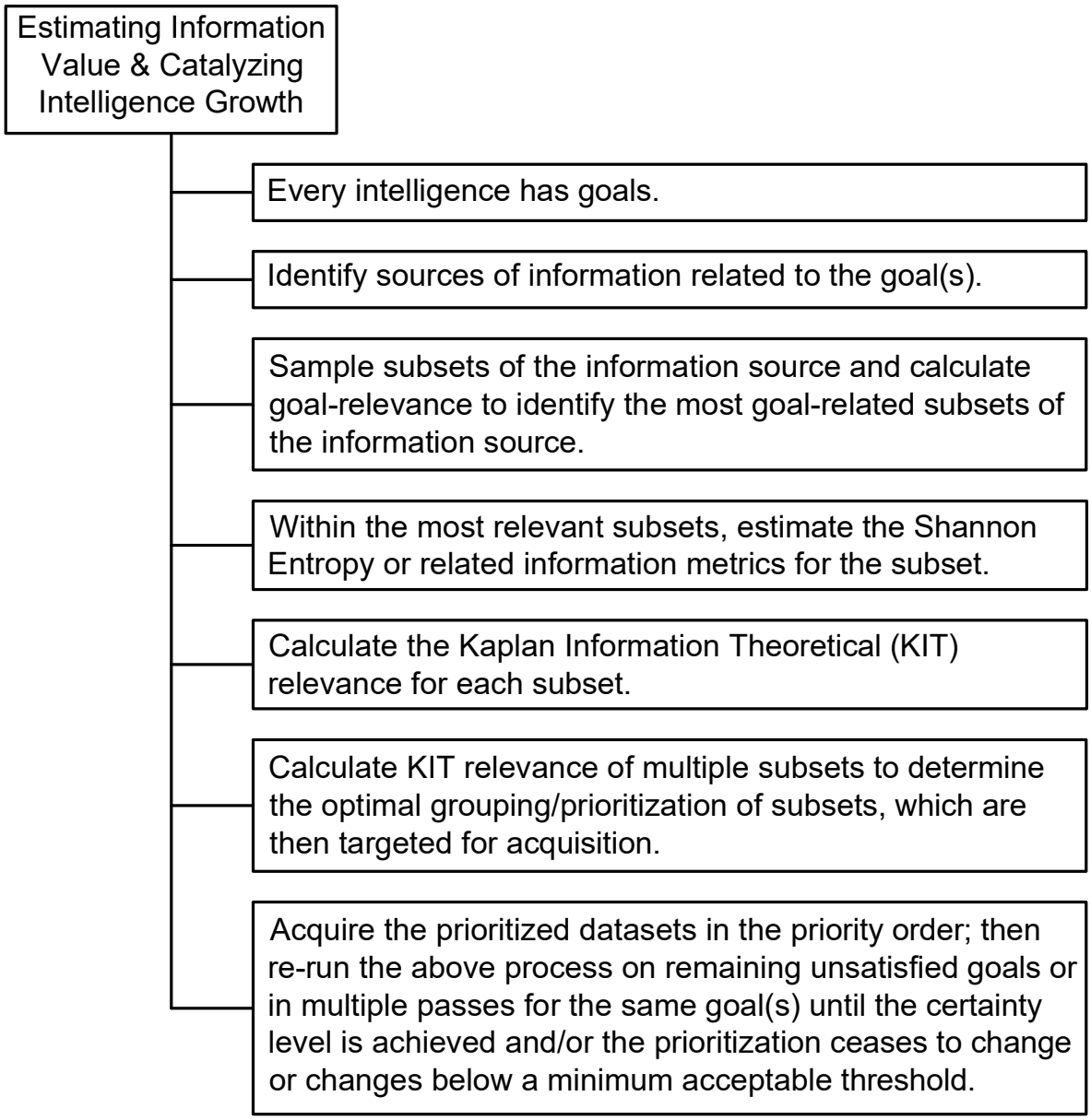
Estimating Information Value & Catalyzing Intelligence Growth
Every intelligence has goals.
Identify sources of information related to the goal(s).
Sample subsets of the information source and calculate goal-relevance to identify the most goal-related subsets of the information source.
Within the most relevant subsets, estimate the Shannon Entropy or related information metrics for the subset.
Calculate the Kaplan Information Theoretical (KIT) relevance for each subset.
Calculate KIT relevance of multiple subsets to determine the optimal grouping/prioritization of subsets, which are then targeted for acquisition.
Acquire the prioritized datasets in the priority order; then re-run the above process on remaining unsatisfied goals or in multiple passes for the same goal(s) until the certainty level is achieved and/or the prioritization ceases to change or changes below a minimum acceptable threshold.

FIG. 21

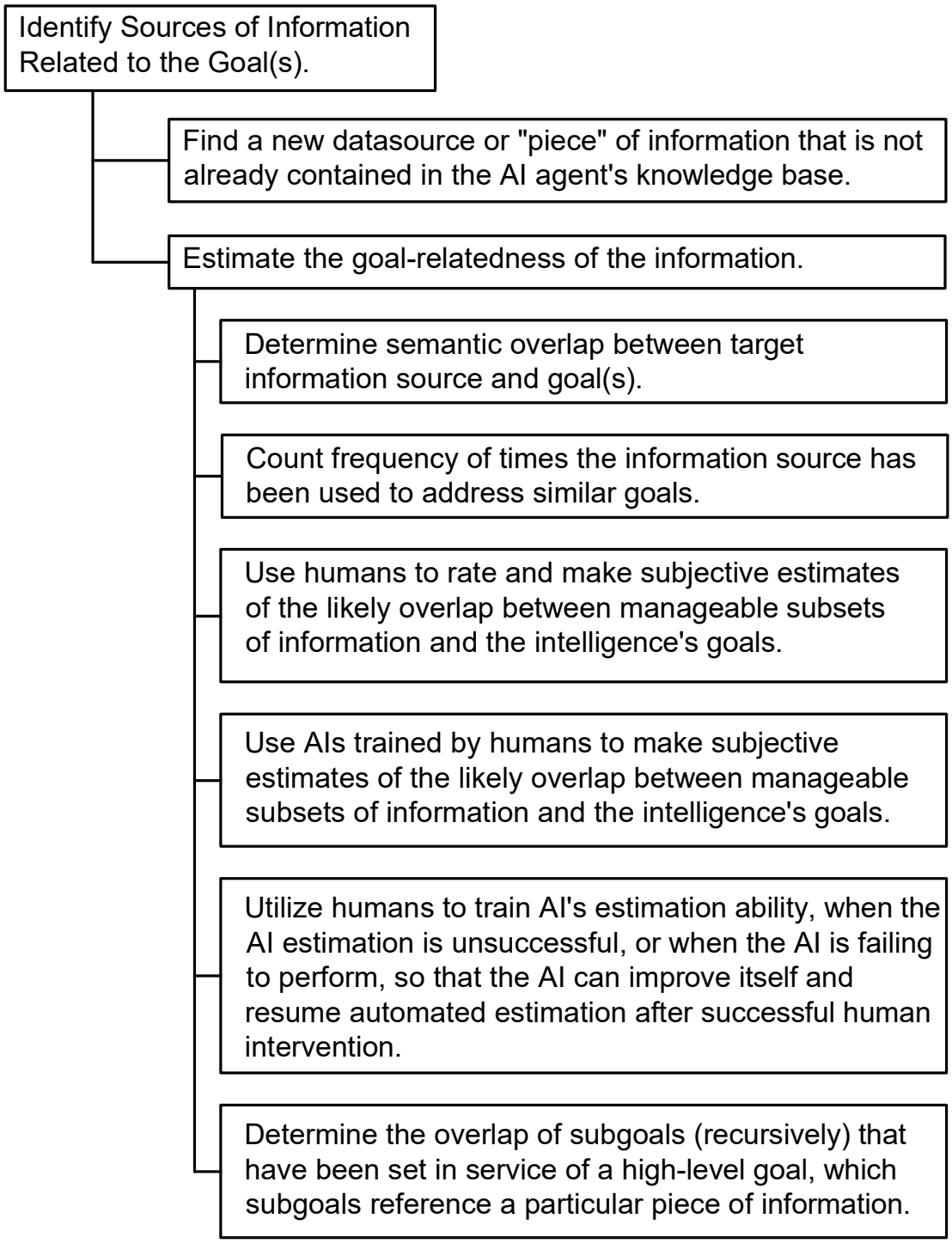
Identify Sources of Information Related to the Goal(s).
Find a new datasource or "piece" of information that is not already contained in the AI agent's knowledge base.
Estimate the goal-relatedness of the information.
Determine semantic overlap between target information source and goal(s).
Count frequency of times the information source has been used to address similar goals.
Use humans to rate and make subjective estimates of the likely overlap between manageable subsets of information and the intelligence's goals.
Use AIs trained by humans to make subjective estimates of the likely overlap between manageable subsets of information and the intelligence's goals.
Utilize humans to train AI's estimation ability, when the AI estimation is unsuccessful, or when the AI is failing to perform, so that the AI can improve itself and resume automated estimation after successful human intervention.
Determine the overlap of subgoals (recursively) that have been set in service of a high-level goal, which subgoals reference a particular piece of information.

FIG. 22

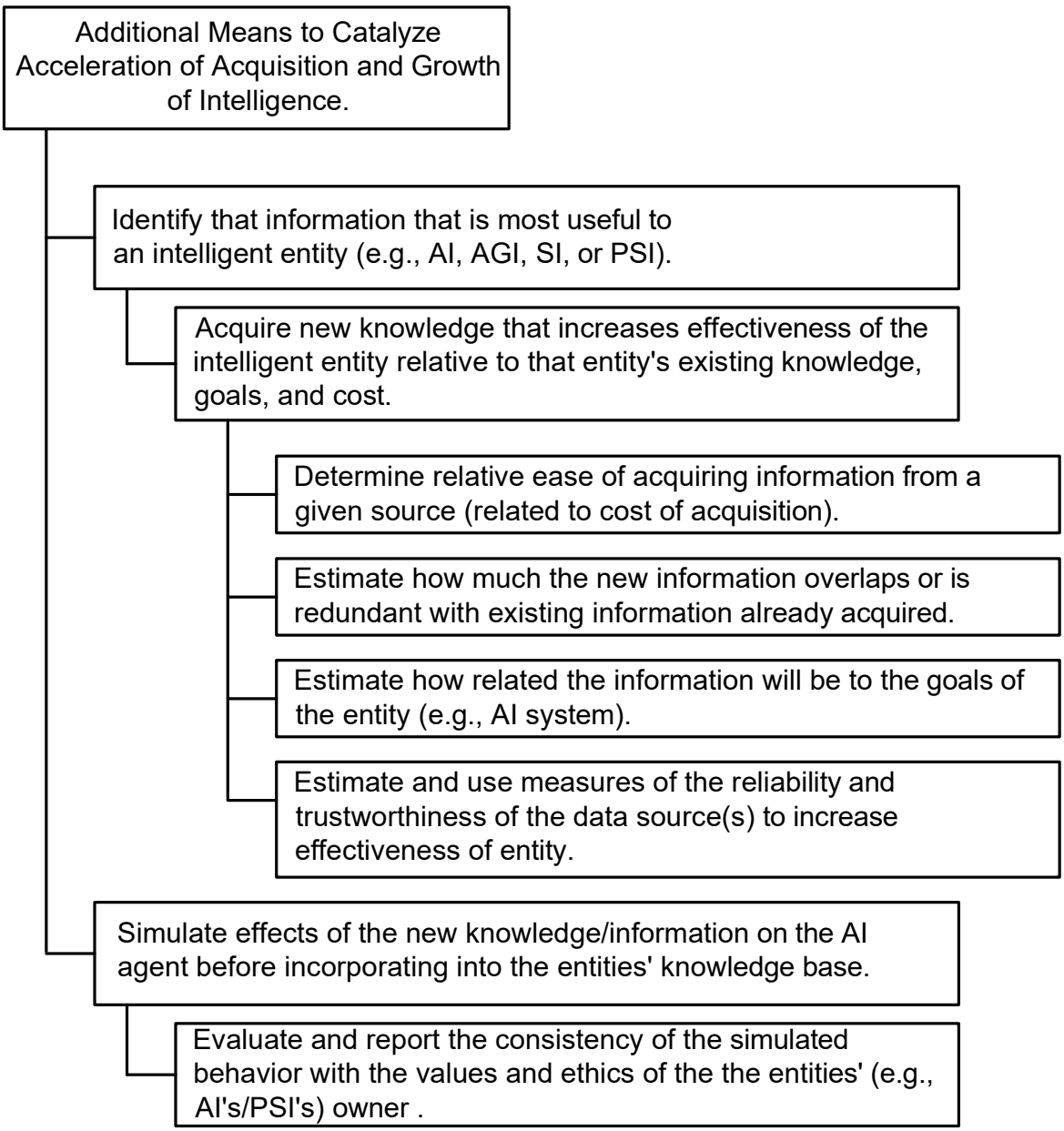
Additional Means to Catalyze Acceleration of Acquisition and Growth of Intelligence.
Identify that information that is most useful to an intelligent entity (e.g., AI, AGI, SI, or PSI).
Acquire new knowledge that increases effectiveness of the intelligent entity relative to that entity's existing knowledge, goals, and cost.
Determine relative ease of acquiring information from a given source (related to cost of acquisition).
Estimate how much the new information overlaps or is redundant with existing information already acquired.
Estimate how related the information will be to the goals of the entity (e.g., AI system).
Estimate and use measures of the reliability and trustworthiness of the data source(s) to increase effectiveness of entity.
Simulate effects of the new knowledge/information on the AI agent before incorporating into the entities' knowledge base.
Evaluate and report the consistency of the simulated behavior with the values and ethics of the the entities' (e.g., AI's/PSI's) owner .

FIG. 23

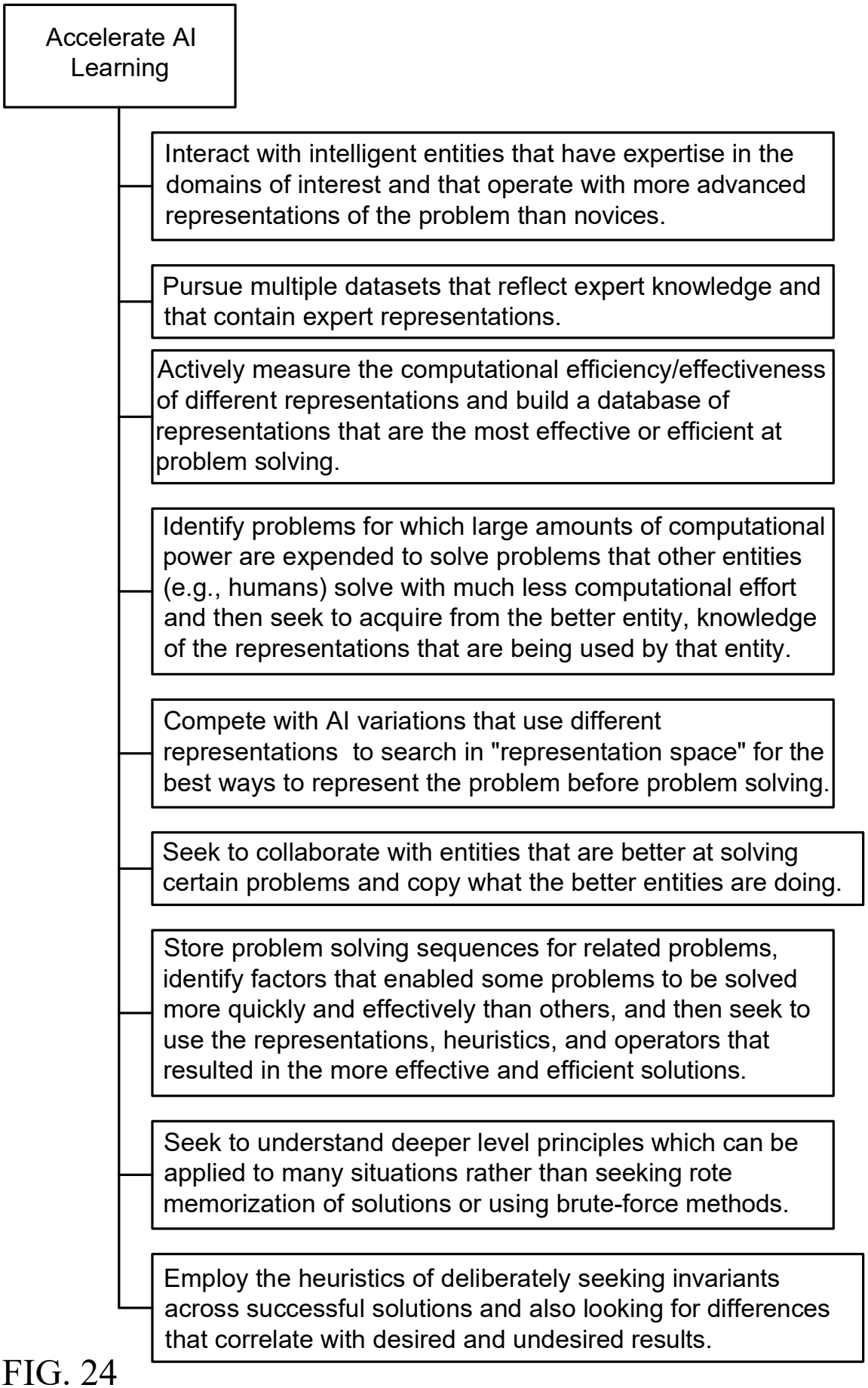
Accelerate AI Learning
Interact with intelligent entities that have expertise in the domains of interest and that operate with more advanced representations of the problem than novices.
Pursue multiple datasets that reflect expert knowledge and that contain expert representations.
Actively measure the computational efficiency/effectiveness of different representations and build a database of representations that are the most effective or efficient at problem solving.
Identify problems for which large amounts of computational power are expended to solve problems that other entities (e.g., humans) solve with much less computational effort and then seek to acquire from the better entity, knowledge of the representations that are being used by that entity.
Compete with AI variations that use different representations to search in "representation space" for the best ways to represent the problem before problem solving.
Seek to collaborate with entities that are better at solving certain problems and copy what the better entities are doing.
Store problem solving sequences for related problems, identify factors that enabled some problems to be solved more quickly and effectively than others, and then seek to use the representations, heuristics, and operators that resulted in the more effective and efficient solutions.
Seek to understand deeper level principles which can be applied to many situations rather than seeking rote memorization of solutions or using brute-force methods.
Employ the heuristics of deliberately seeking invariants across successful solutions and also looking for differences that correlate with desired and undesired results.

FIG. 24

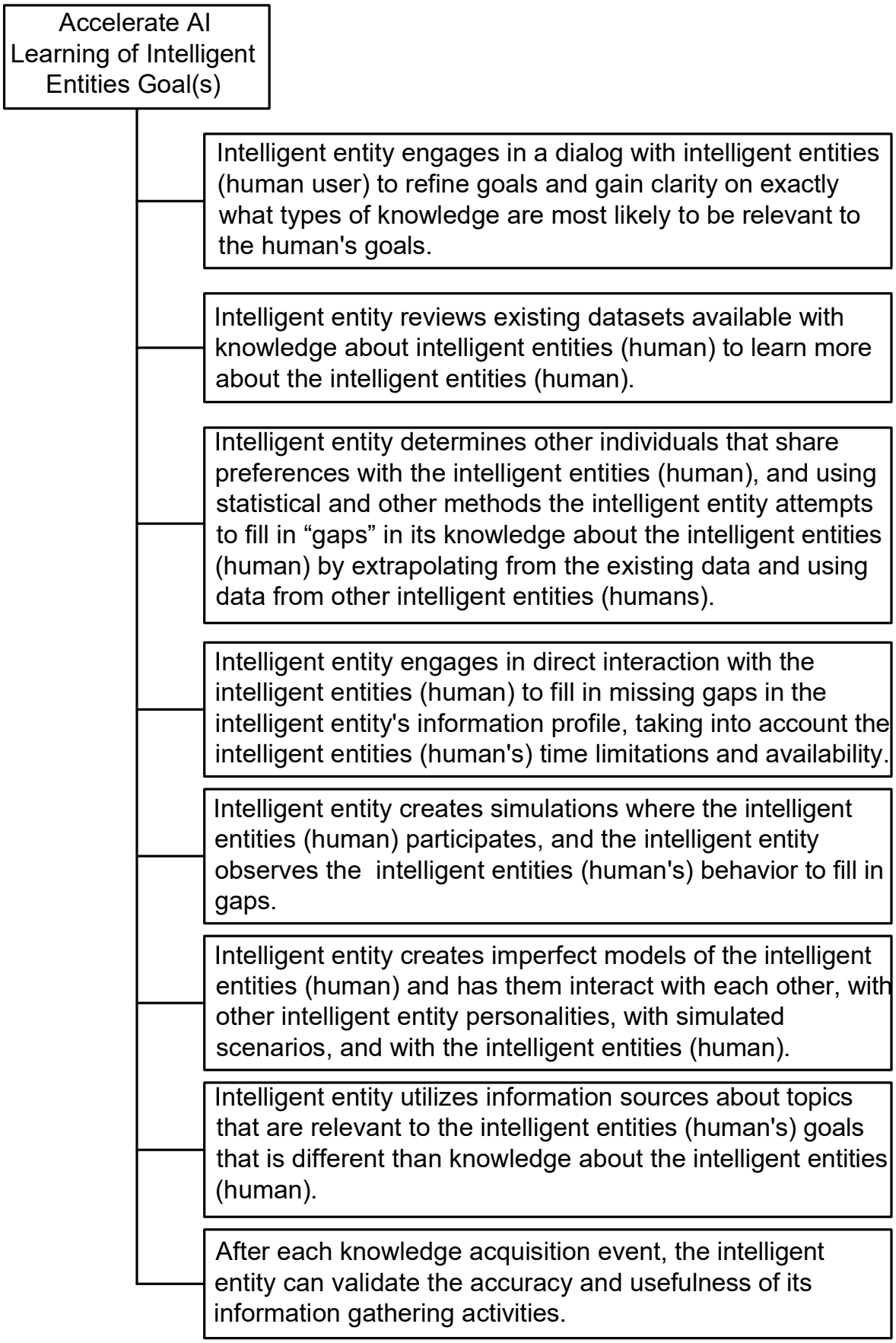
Accelerate AI Learning of Intelligent Entities Goal(s)
Intelligent entity engages in a dialog with intelligent entities (human user) to refine goals and gain clarity on exactly what types of knowledge are most likely to be relevant to the human's goals.
Intelligent entity reviews existing datasets available with knowledge about intelligent entities (human) to learn more about the intelligent entities (human).
Intelligent entity determines other individuals that share preferences with the intelligent entities (human), and using statistical and other methods the intelligent entity attempts to fill in “gaps” in its knowledge about the intelligent entities (human) by extrapolating from the existing data and using data from other intelligent entities (humans).
Intelligent entity engages in direct interaction with the intelligent entities (human) to fill in missing gaps in the intelligent entity's information profile, taking into account the intelligent entities (human's) time limitations and availability.
Intelligent entity creates simulations where the intelligent entities (human) participates, and the intelligent entity observes the intelligent entities (human's) behavior to fill in gaps.
Intelligent entity creates imperfect models of the intelligent entities (human) and has them interact with each other, with other intelligent entity personalities, with simulated scenarios, and with the intelligent entities (human).
Intelligent entity utilizes information sources about topics that are relevant to the intelligent entities (human's) goals that is different than knowledge about the intelligent entities (human).
After each knowledge acquisition event, the intelligent entity can validate the accuracy and usefulness of its information gathering activities.

FIG. 25

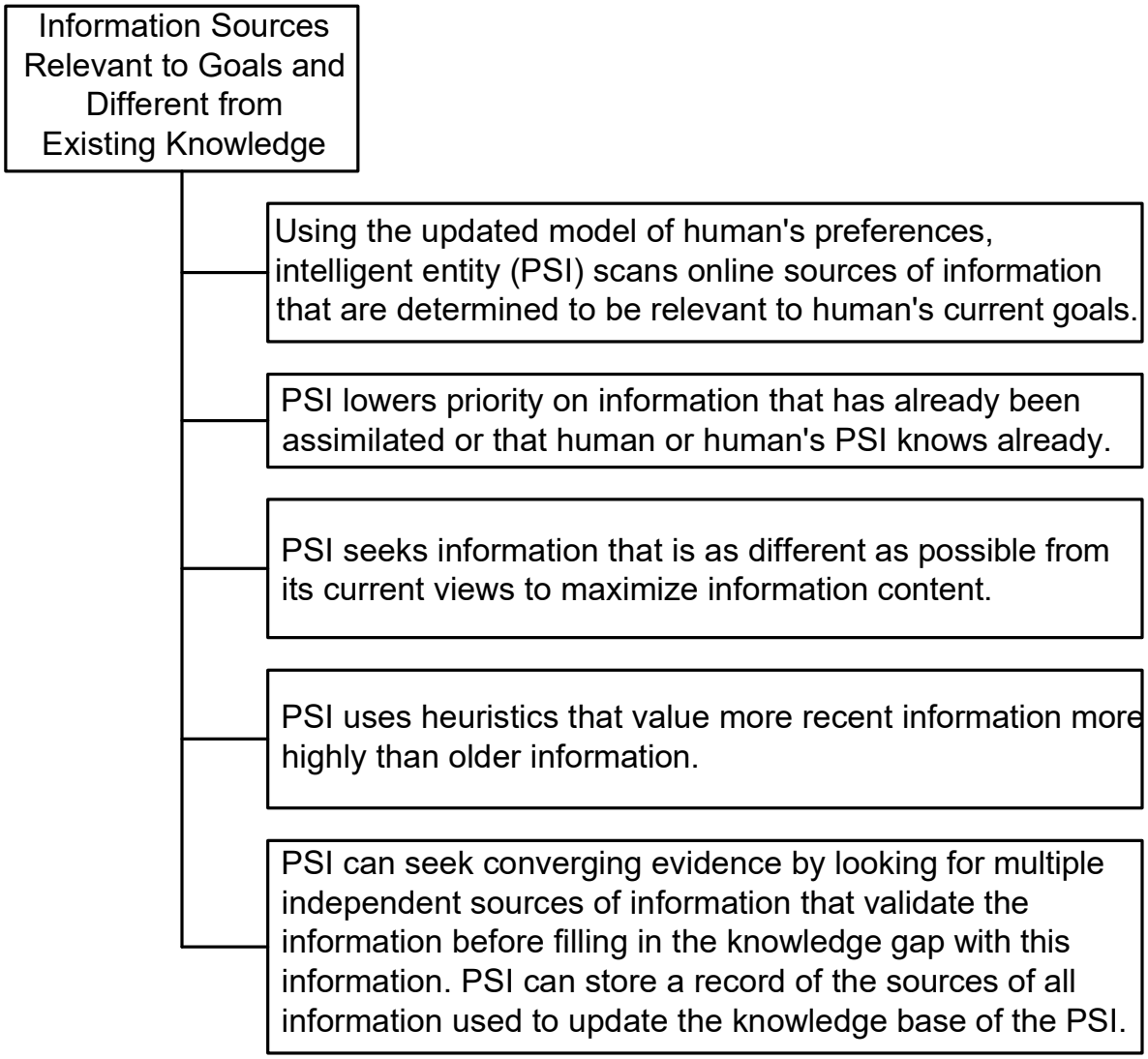
Information Sources Relevant to Goals and Different from Existing Knowledge
Using the updated model of human's preferences, intelligent entity (PSI) scans online sources of information that are determined to be relevant to human's current goals.
PSI lowers priority on information that has already been assimilated or that human or human's PSI knows already.
PSI seeks information that is as different as possible from its current views to maximize information content.
PSI uses heuristics that value more recent information more highly than older information.
PSI can seek converging evidence by looking for multiple independent sources of information that validate the information before filling in the knowledge gap with this information. PSI can store a record of the sources of all information used to update the knowledge base of the PSI.

FIG. 26

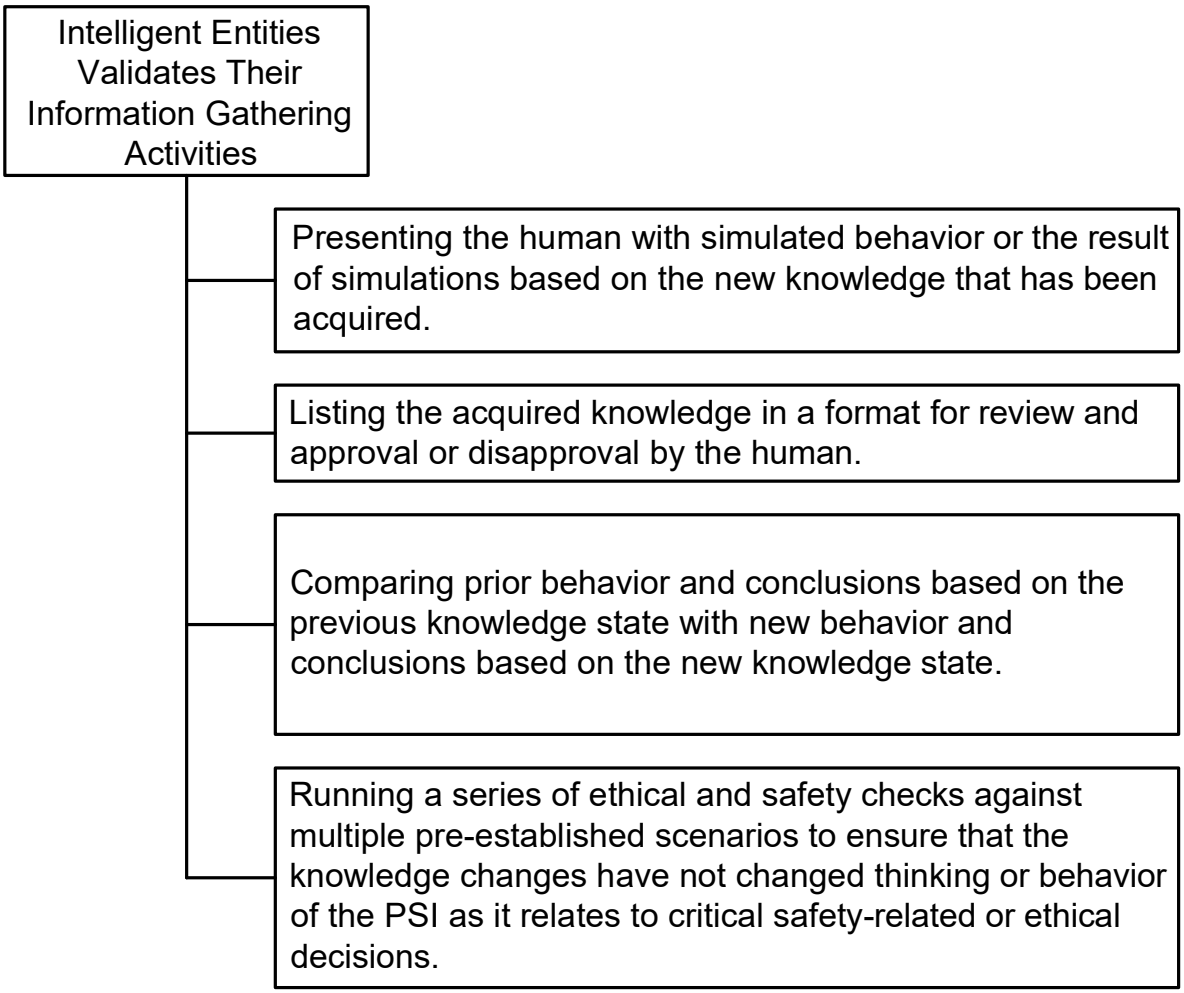
Intelligent Entities Validates Their Information Gathering Activities
Presenting the human with simulated behavior or the result of simulations based on the new knowledge that has been acquired.
Listing the acquired knowledge in a format for review and approval or disapproval by the human.
Comparing prior behavior and conclusions based on the previous knowledge state with new behavior and conclusions based on the new knowledge state.
Running a series of ethical and safety checks against multiple pre-established scenarios to ensure that the knowledge changes have not changed thinking or behavior of the PSI as it relates to critical safety-related or ethical decisions.

FIG. 27

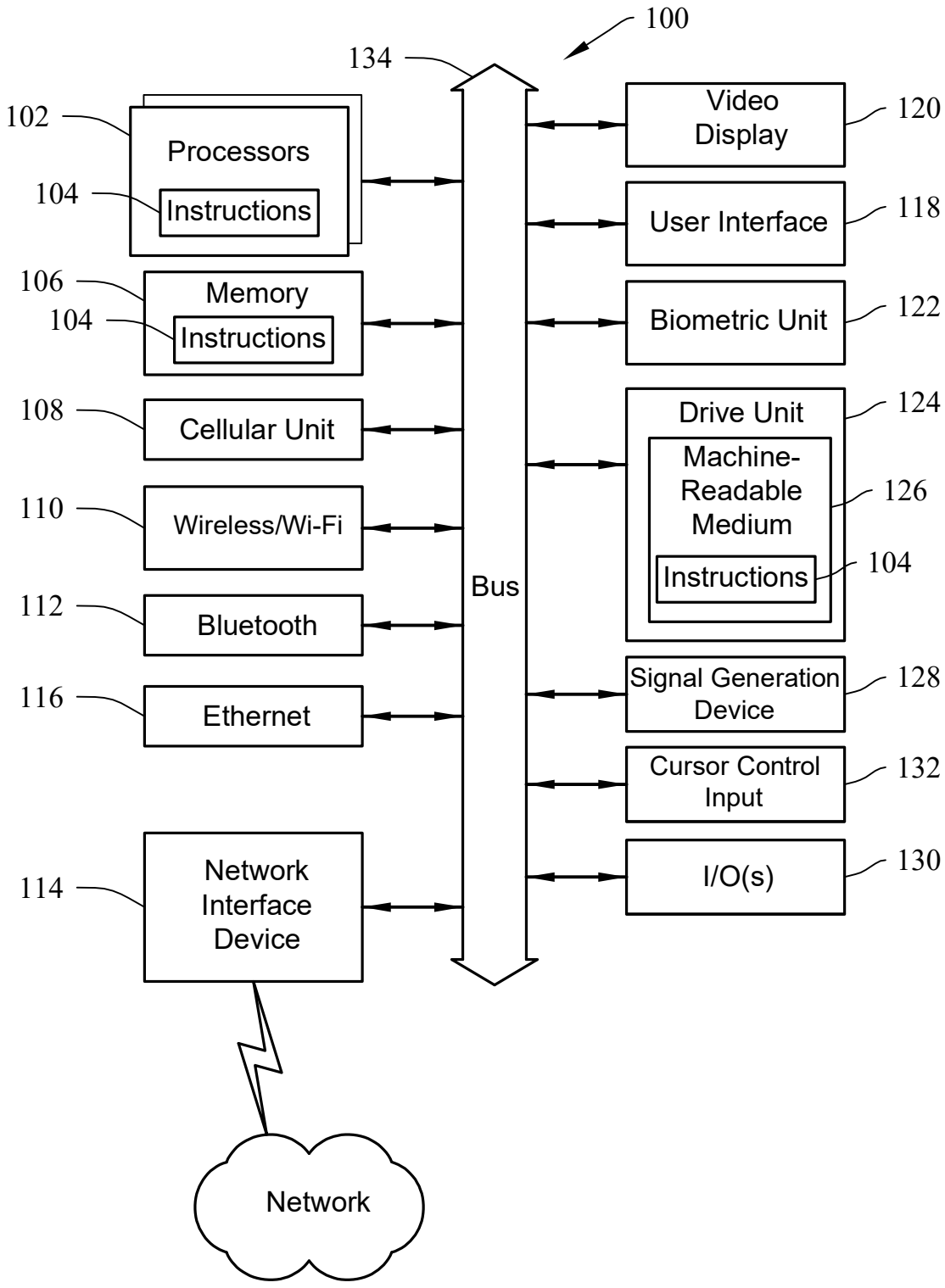
100
134
102
Processors
104
Instructions
106
Memory
104
Instructions
108
Cellular Unit
110
Wireless/Wi-Fi
112
Bluetooth
116
Ethernet
114
Network Interface Device
Bus
Video Display
120
User Interface
118
Biometric Unit
122
Drive Unit
124
Machine-Readable Medium
126
Instructions
104
Signal Generation Device
128
Cursor Control Input
132
I/O(s)
130
Network

FIG. 28