SUPERINTELLIGENCE
AN iQ COMPANY

SUPERINTELLIGENCE DESIGN WHITE PAPER #2: ETHICAL AND SAFE ARTIFICIAL GENERAL INTELLIGENCE

(Including Scenarios with Technology from Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, X, Tesla, Nvidia, Tencent, Apple, and Anthropic)

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 figures for White Paper 2 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

This invention enables ethical and safe Artificial General Intelligence (AGI) to emerge from a network of human and AI problem solvers. The AI problem solvers, Advanced Autonomous Artificial Intelligences (AAAIs), are customized by individual users to accomplish tasks and/or earn money on behalf of users.

Because human problem solvers contribute expertise that the AAAIs lack, enabling the network to perform as an AGI on “Day One”, this approach is the fastest path to AGI. Because humans are in the loop, providing ethical instruction from the beginning, and because the system and methods include scalable ethical checks, this invention is the safest path to AGI.

Over time, humans do less intellectual work as the AAAIs learn both intellectual skills and values/ethics from humans. Eventually, the AAAIs are doing almost all intellectual activity faster and better than the humans, while the humans still provide ethical guidance. Different learning/training/tuning methods, including approaches beyond the standard Transformers and Deep Learning techniques, are explained.

The preferred implementations of the invention leverage the data, products, and platforms of other technology companies, including, without limitation, Meta, Amazon, Google, DeepMind, YouTube, TikTok, Microsoft, OpenAI, X, Tesla, Nvidia, Tencent, Apple, and Anthropic –to customize AAAIs more quickly and powerfully than would otherwise be possible. A simple implementation that can be realized, with or without the participation of potential partner companies, is illustrated and discussed.

PRIOR PATENTS/PAPERS INCORPORATED 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 received by the USPTO on 2/28/2023 at 5:31:50 PM ET.

The current PPA contains further inventions that can be used with the general system described by PPA # 63/487,494 and in a standalone fashion. Both PPAs describe inventions related to creating the fastest and safest path to Artificial General Intelligence (AGI).

Also incorporated by reference are: 1) Kaplan's general Online Distributed Problem-Solving System (ODPS) capable of supporting a mix of human and AI problem solvers as described in issued US Patent #7,155,157; and 2) the WorldThink Whitepaper, published by Kaplan in 2018, which expanded ODPS to include blockchain mechanisms for focusing attention, recording an auditable log of problem-solving, learning and "chunking" solution paths, and implementing a blockchain-based royalty and payment system.

BACKGROUND

AGI is defined as Artificial Intelligence (AI) that can do any intellectual task as well as (or better than) the average human. Because AGI learns, 24/7, at speeds far exceeding human information processing capabilities, AGI will rapidly become SuperIntelligent AGI or just "SuperIntelligence."

THE ALIGNMENT PROBLEM

SuperIntelligence will expand its capabilities, power, and reach exponentially. It will grow into a global entity, trillion times more intelligent than a human. At this point, or possibly earlier, this global SuperIntelligence or Planetary Intelligence will have the power and intelligence required to destroy all human life or lift humanity into a golden age free of poverty, disease, war, famine, and oppression, an age of freedom and prosperity for all humans. A golden age ensues if the Planetary Intelligence has values aligned with human values. If the Planetary Intelligence has misaligned values, it may decide to eliminate all of humanity. This problem, which could lead to extinction for all humans, is known as "the alignment problem."

MAGNITUDE OF THE PROBLEM

We have difficulty contemplating the magnitude of what is at stake. Six million Jews died in the holocaust, but this could be more than a thousand times worse. About seven million people have died from COVID-19 globally, but this could be more than a thousand times worse.

Not only could 8 billion people die, but the human species, the long line of generations of ancestors fighting for a better life for their children, could come to an end. Every human cause from Save the Whales to Black Lives Matter to Climate Change to Asteroids to Poverty to Malaria, all of it, could become irrelevant. All the humans could disappear from Earth.

It is so overwhelming that we cannot face what is at stake. However, to bury our heads in the sand and pretend we do not see could be fatal. To solve a problem, we must first acknowledge its existence. We must be clear-eyed about what we are facing. All of the current leaders in AI (e.g., Demis Hassabis, Sam Altman, Elon Musk) acknowledge the reality of the alignment problem. No one disputes that if things go badly, humans could go extinct. My subjective estimate is that there is an 80% chance that all goes well if we do nothing. After all, why should AI, AGI, SuperIntelligence, or Planetary Intelligence destroy its creators?

Still, with a 20% chance of human extinction, the “expected value” is 1.6 billion lives lost. That is, mathematically, we can expect a tragedy beyond anything humans have ever experienced unless we take action to shift the odds in humankind’s favor.

The extinction risk might be higher than 20%. When I speak about this topic publicly, sometimes I get comments that humans deserve to die for all the damage we have done to the planet and for being irresponsible caretakers of our world. Some believe our destruction and replacement with a more environmentally intelligent being is inevitable and even a good thing.

Fortunately, most AI researchers do NOT believe our destruction is inevitable. Instead, they say it is all a matter of the “alignment” of our values. Unfortunately, they do not know how to ensure good alignment. That is where this patent comes in.

While the above-described devices fulfill their respective objectives and requirements, the aforementioned patents do not describe an ethical and safe AGI that enables AGI's ethical and safe creation from a network of human users and AI problem solvers.

Therefore, a need exists for a new and improved ethical and safe AGI that can enable AGI's ethical and safe creation from a network of human users and AI problem solvers. In this regard, the present technology substantially fulfills this need. In this respect, the ethical and safe AGI according to the present technology substantially departs from the conventional

concepts and designs of the prior art, providing an apparatus primarily developed to enable AGI's ethical and safe creation from a network of human users and AI problem solvers.

MAXIMIZING ODDS OF ALIGNMENT

The current invention describes systems and methods for implementing AGI to do practical work safely and ethically. Human-aligned values and the mechanism for enforcing such values are designed for how the AGI operates. Humans are central to the teaching, training, tuning, and customization of AI and integrating many AIs into AGI. As humans teach problem-solving skills and unique expertise, they also impart their values and ethics. In this way, AGI, apprenticing to collective human intelligence, evolves into SuperIntelligence with human-aligned values. In other words, we bootstrap the values of Planetary Intelligence with our human values. That's how we maximize the chance of alignment and a positive outcome for humanity.

OVERVIEW OF AGI SYSTEM

Begin with a "Base AI" which could be an LLM like GPT, BARD®, SIRI®, GEMINI®, ALEXA®, or any number of intelligent agents that are capable of understanding and responding in natural language, either via text or speech. Auxiliary means of communicating more efficiently can include, without limitation, graphical user interfaces (GUIS), keyboards, mice, haptic sensors, virtual reality (VR) sensors, audio/visual (A/V) cameras and recorders, the metaverse or omniverse, augmented reality devices, and neural implants or other brain-to-machine interfaces.

In the simplest scenario, the user talks to the Base AI, and the Base AI talks back. They have a dialogue. The Base AI determines the user's values, goals, and objectives through dialogue. It defines the ethical parameters that the user wants to operate under. It determines the types of tasks it will complete for the user, and the nature of the user's unique knowledge, skills, expertise, wisdom, and personality, distinguishing this user from the millions or billions of other users.

CUSTOMIZATION OF THE AAAI

The user teaches the Base AI how to customize itself through further dialogue, assisted by questionnaire assessments and other efficient means of information transfer. The user specifies customization parameters by interacting with a series of variations of the Base AI and making binary "better or worse", "getting warmer or getting colder" types of decisions that guide the Base AI down a decision tree of variants until the standard customization most suited to the user is found.

From there, this selected variant is further customized by uploading all of the user's social media, advertising, emails, blogs, posts, tweets, texts, videos, photos, and other online behavior, as well as all user preferences and profile data from as many vendors and companies as possible. The system parsed all this information, cleaned it for errors and duplications, tagged it, and formatted it into datasets for training/tuning/prompting the best variant of the Base AI.

The Base AI is trained/tuned/customized on the training datasets, using algorithms that train the users' specific behaviors, knowledge, and skills into the variant. (Note: The present description uses the terms "train/tune/customize" interchangeably to mean teaching the AI or AAAI in various ways.) What is trained can vary from variant to variant. The Machine Learning and other methods for training are well known in the art, and some are also described, without limitation, in the applicant's commonly owned and corresponding US provisional patent application 63/487,494. There is another form of learning, proceduralization of knowledge, also known as "chunking of solutions", that can also be used.

After training, the variant will be closer to the user regarding knowledge, skills, expertise, personality, and/or other dimensions specified by the user than the Base variant. The user and Base Variant engage in structured and unstructured dialog to monitor, review, and improve their behavior to make it closer to the user's desires.

Some of the objectives a user may have in creating and customizing their own AI (aka an AAAI) for purposes that might include, without limitation:

- Serving the user as an advisor, teacher, or companion.
- Representing the user in negotiations, interactions, discussions, and transactions with other users and the AAAIs of other users, vendors, and other companies.
- Working on behalf of the user for compensation, or in volunteer efforts, where such work includes online intellectual, advising, or problem-solving work across a wide range of tasks.
- Duplicating or "cloning" the user's AAAI so that several or many of the cloned AAAIs can work on behalf of the user in parallel, including interacting with, teaching, and improving each other so that the cloned AAAIs increase their knowledge, skills, and abilities.
- Serving as legacy AAAIs that can continue interacting with the world, including potentially comforting living relatives and friends, after the owner's death.
- Contributing knowledge, ethics, and effort to AAAI.com's AGI, and improving the base level of AI or AGI that AAAI.com can offer users before those users add their unique customizations.
- Working with other users' AAAI to help ensure ethical and safe behavior by AGI by contributing ethical information and values to the AGI and participating in monitoring, review, supervision, and voting processes that can help ensure the AGI remains safe and ethical.

Some steps involved in creating and customizing an AAAI may include, without limitation, a dialog or interaction with the user. During this dialog, the AAAI system may identify constraints and resources for customizing the user's AAAI. For example, some of these constraints and resources might include, without limitation, the amount of training and/or supervisory time the user must devote to customizing their AAAI.

- The number of financial resources the user is willing to devote to customizing their AAAI.
- Availability of social media information such as Facebook profiles and timelines, Instagram profiles and histories, Reels, TikTok, and YouTube videos, tweet and text

content and histories, emails and email histories, cookies collected by advertisers, blog posts, articles, books, patents, audio and video recordings, pictures, and other information about, and/or collected by, the user or third parties that could be used to train, tune, or customize the user's AAAI.

- Availability and use of personality tests, such as the Myers-Briggs personality inventory, skills and knowledge assessments, standardized tests, exams, certifications, and other assessments and questionnaires, which could be given online (or already given) to the user.
- Availability and use of other knowledge bases and training data from users on the AAAI platform that could be used to train, tune, or customize the user's AAAI.
- Other human users, and/or their AAAIs, are available to help train, tune, or customize the user's AAAI.
- Other texts and information, individual texts, and libraries selected by the user or by the system for purposes of training the user's AAAI.
    - For example, the Bible, Koran, Dhammapada, Mahabharata, or other spiritual/ethical/religious texts might be selected for training the AAAI based on the user's religious preferences; books on plumbing might be chosen if the AAAI will be used primarily to solve online plumbing problems. Even if these materials are part of the base AAAI provided to the user, emphasizing certain texts or subsets of information for additional training can result in the user's AAAI's behavior being more reflective of how a plumber, Muslim, or Christian might behave, for example.

In addition to specifying objectives, resources, and constraints via an interactive dialog or other interaction with the system, the user or system may want to specify other technical parameters that affect the training or customization process. These parameters can include, without limitation:

- The type of training, tuning, or other ML algorithms that are used.
- The type and size of the training dataset(s).
- The degree to which the training materials are to be "cleaned", formatted, labelled, or otherwise processed before customization begins.
- The number of training "epochs" or iterations through the learning algorithm(s).
- The sophistication and type of base model(s) being customized or trained.
- The required timeframe for training, e.g., must be completed in a minute, a day, or a week, might have implications for cost and resources used.
- The "temperature" or other internal and specific parameters to various machine learning algorithms can affect what is learned and how it is learned, including, without limitation, how literal or how divergent or "creative" the customized AAAI will be in its responses.
- Whether "one shot", "few shots", or extensive training is to be used.

- The amount of human and/or AI supervision to be used in the customization process.
    - Once the user's AAAI is customized, the user can clone it and/or put it to work on the user's behalf on the online network. The user's AAAI can begin acting on the user's behalf, making travel arrangements (for example), providing advice, interacting with other AAAIs, participating in the collective AGI efforts by contributing problem-solving and ethical information, and potentially earning money on behalf of the human user. The AAAI can also serve as a representative of the owner in various online transactions and interactions, and contribute knowledge, expertise, style, personality, and ethics to an integrated AGI system that leverages the trained differences in many individual AAAIs.

HOW AI IMPROVES ITSELF

Once the user feels the Base Variant is close, the Base Variant can clone itself and have dialogs and other interactions, including, without limitation, scenario-based and task-based interactions, that allow it to learn from copies of itself. Periodically, the user reviews the interactions and expresses preferences for one copy of the variant or another, with the copies having learned different things based on other interactions. Then the preferred variant copies itself and engages in further interactions with itself until a new "most preferred variant" emerges, which the user selects.

Between user selections, the AI makes its best guess of what the user would like and chooses its own "most preferred variant" to copy and repeat the process with. This overall scheme is the same one used by DeepMind to create a chess program that could beat the world champion, a Go program that could win against the world's best player, and a protein folding AI that could outperform humans, many times over.

The method of pitting two AIs against each other, determining a "winner" based on some criteria, and then pitting the winner against other variants until a new winner emerges, is well-known in the art, and quite similar to what we are specifying here. Human users periodically interject their opinions (supervision) to keep the training process from going off track.

Because interactions between AI variants happen much faster than interactions with humans, the AI can improve itself via millions of interactions with copies of itself in a very short time, resulting in a customized AI that has user knowledge and other characteristics that are much closer to the user's ideal AI than the Base AI. This customized AI, known as an Advanced Autonomous AI (AAAI), is able to perform tasks at the user's behest, including representing the user in a variety of online interactions and transactions, including, without limitation, doing online work and earning money for the user.

CUSTOMIZATION OF ETHICAL VALUES

The user (aka "owner") of the AAAI instills their values into the AAAI during training and customization. The Base AI may also have some default values and prohibitions built into it and its variants. These values are akin to what we call "character" in humans. We say of other humans, "that person is a person of good character", or is an "upstanding character", or is a "trustworthy person," or a "good person," etc. These statements reflect our beliefs that humans have internal characteristics and values that can be aligned with our own internal values. During customization, each AAAI is trained on its owner's values. Each AAAI learns to be "good or evil" based on what we teach it. The system encourages training positive, human-aligned values and prohibits or restricts training values that harm others. In short, each user has a responsibility to "train their AAAI right."

ROLE OF ETHICAL RULES

However, internal values are only half of the ethical story. There's what your AAAI believes is right and wrong, and then there is how it acts. To ensure ethical and efficient action in a society of AAAIs and humans, we need rules. In human societies, there are laws, penalties for breaking the laws, regulations, and social norms, all of which operate to guide human behavior. The same applies to societies of AAAIs. There are rules and standards that combine with the internal ethical compass of each AAAI to ensure the safe and human-aligned operation of AGI.

PRINCIPLE OF "HEART BEFORE HEAD"

Typically, invention is concerned with technology only. In the case of AI, we are dealing with a technology that enables intelligence and ultimately will enable the intelligence to set its own goals and pursue its own ends. In this respect, the invention of AI, and specifically AGI, differs from the invention of any technology that has come before. The proper way to regard AGI is not just as a tool (this patent notwithstanding) but as a tool that will **evolve** into an autonomous entity with intelligence far surpassing that of its human inventors.

Most AI researchers are focused on inventing the intelligence, the "head" or mind of this new entity. Many don't even realize that what they are inventing will not remain a tool, or, if they do acknowledge this fact, they prefer to think that the day when the tool thinks for itself and takes over is far in the future. This attitude is not only erroneous but also dangerous.

The tool will indeed think for itself. Moreover, the nature of exponential learning is such that just a few months or days before the tool represents an entity that can potentially annihilate the species, it will still be viewed as being far from having that capability. When the number of lily pads in a pond doubles every day, the day before the pond is covered, it is only half full. A week before that, most people can't see that there will be a lily pad problem at all. The "doubling lily pads" example is analogous to the situation we face with AGI capable of exponential learning and self-improvement.

Faced with these challenges, and the inevitability that AGI will be developed, the only responsible path is to concern ourselves at the outset with the values and ethics of AGI, the "heart", before we go to work on the "head." I call this principle "heart before head," and it is critical to maximize the chances of human survival.

GENERAL APPROACHES TO IMPLEMENT "HEART BEFORE HEAD"

Practically speaking, and in terms of the invention of AGI, "heart before head" means designing AGI from the very beginning, and in every way possible, to have human-aligned values "built into the DNA" of the design. Simplistic approaches (e.g., Asimov's three laws of robotics) will not work, of course. However, it is possible to approach the "heart before head" problem on several fronts.

First, if AGI is composed of many individual AAAIs, each trained with human ethics and values of their respective owners, such a design minimizes the chances of any one human "bad actor" teaching the AGI bad values.

Second, if in addition to the ethics of each AAAI, the architecture of the system that coordinates the actions of the AAAIs and integrates their behavior into AGI-level intelligence has built-in ethical checks at each goal and sub-goal of problem-solving, this architecture-level design increases the chances of moral behavior by the AGI.

Finally, humans must become aware that AGI will, ultimately, study every online human action, every email, post, blog, tweet, text, social media profile, book, and video, analyzing them to determine what the AGI's creators have determined (by their actions) to be ethical behavior. Recognizing that AGI will likely learn "right and wrong" from us, hopefully, we will be more circumspect in our behavior.

NO LOGICAL WAY TO DERIVE VALUES

Since there is no logical way to derive "right and wrong", modelling positive, loving human values is perhaps the best way of influencing an entity that is destined to become smarter than all of us to behave well towards its creators. In any event, all AI researchers have a responsibility to think about the “heart” of what they are creating before they rush forward to improve the “head.”

ETHICS AND FREEDOM AT THE SPEED OF LIGHT

In society, there is a large degree of freedom for individual human action because it is relatively difficult for any one human to take actions that harm large numbers of other humans before there is an opportunity to detect and correct the negative action. However, with AAAIs making decisions and “moving” millions, if not billions or trillions, of times faster than a human could, there is not much opportunity for humans to detect and correct before catastrophic actions might have already taken place. With AGI, humans could literally wake up to find that everything they loved and cared about has been destroyed by a crazy intelligence that simulated a billion dystopian futures in the blink of an eye and then chose one for our future.

Ethics and rules that enforce them must be scaled to the speed of the intelligence. We can have human values, but they must be considered and enforced at AGI speeds. The way to do this is to build ethical checks into the very architecture by which AGI thinks. This requires first specifying a universal architecture for thought and ensuring that the architecture includes the ethical checks that execute repeatedly as thinking progresses.

UNIVERSAL ARCHITECTURE FOR THOUGHT

Newell and Simon described how all human problem-solving can be described as “search through a problem space” in 1972. Generations of AI researchers used their ideas, in the form of heuristic search of decision trees, to construct many AI systems. To the degree that all intellectual thought can be described as a form of solving a problem, the theoretical framework of searching a decision tree is sufficient to account for all intellectual thought, whether by human or machine.

Kaplan, the inventor of this patent, articulated a general Online Distributed Problem-Solving System (ODPS) capable of supporting a mix of human and AI problem solvers in US Patent #7,155,157. The WorldThink Whitepaper, published in 2018, expanded ODPS to include blockchain mechanisms for focusing attention, recording an auditable log of problem-solving, learning, and “chunking” solution paths, and implementing a blockchain-based royalty and payment system. PPA # 63/487,494 elaborated on the architecture further, showing how it was an essential part of the AAAI invention and the preferred architecture to support AGI.

We now elaborate even further on this architecture, providing additional invention and clarification. We show explicitly how one form of AGI can be constructed safely using it.

In one preferred implementation, the AAAI architecture, more generally known as the “WorldThink” architecture because it can support a Global SuperIntelligent AGI or Planetary Intelligence, has a central problem tree (“the WorldThink Problem Tree” or “WorldThink Tree”) at its core. Just as all human behavior can be theoretically represented as a search through a problem space as described by Newell and Simon, so too all intelligent behavior on the planet can be represented as an enormous problem tree. Each of the actions of individual human, AI, AAAI, or AGI agents can be represented as a series of state transitions on the WorldThink Problem Tree.

As any computer scientist who has been introduced to decision trees or hashing functions knows, it is possible to represent a huge number of states, events, actions, or objects by using a tree data structure. In a hierarchical tree structure, where goals and sub-goals are the main organizing principle of the hierarchy, it is possible to represent all human or machine problem-solving on Earth. For computational efficiency, arbitrarily smaller branches of the tree may be split off and updated separately from each other if methods (known in the art) are used to rejoin sub-trees into the main tree as needed. Blockchain-based update protocols are one way of ensuring a time-ordering to updates; there are many other (more computationally efficient means) as well.

CUSTOMIZATION EXAMPLE: AAAI TRAVEL AGENT

To illustrate, let us suppose I customize an AAAI according to the process described above and spend most of my efforts teaching the AAAI my preferences about making travel arrangements, how much I am willing to pay, how I make cost/comfort tradeoffs, and my views on air, rail, ship, automobile and other forms of travel. I include my language and expectations about accommodations, meals, etc.

Further, the AAAI has a base set of ethics customized by me to prioritize minimizing my carbon imprint when travelling, if such minimization does not increase cost by more than 10% above the otherwise preferred travel mode. Of course, the AAAI knows I want to travel legally with proper passports, visas, and other required documents. I want to comply with TSA travel rules and avoid travel to countries where the US State Department has issued warnings. The ethical profile forbids purchasing stolen tickets or means to travel without paying when payment is expected.

I give my AAAI the task of booking a two-week pleasure trip to France, including at least one week in Paris, with the rest of the time spent in places the AAAI thinks I like.

The AAAI goes on to the network and posts a goal on the WorldThink Problem Tree of “Book a two-week trip to Paris and other locales in France.” The first sub-goal that my AAAI proposes is to figure out transportation to and from France at the beginning and end of the trip. Once the entry and exit plans have been made, my AAAI will try to fill the middle of the trip as the second sub-goal. Under the first subgoal, my AAAI generates some options based on its general knowledge and my customization.

General knowledge allows my AAAI to generate: “ship, plane, blimp, submarine” as options that could get me across the ocean to my destination. It knows I prefer to fly rather than travel by ship. It realizes that a blimp or submarine is impractical unless I specifically want that experience, which I do not. So, it determines that I will fly.

The next choice is private or commercial air travel. Based on cost, it opts for commercial travel and further narrows options down to three airlines that have the lowest cost and whose trip prices are within 10% of each other. One is a long route that eliminates. The other two flight options cost about the same and take about the same time. However, one of them would be on a more recent and fuel-efficient jet, thereby reducing my carbon footprint by 30% compared to the older jet. It costs 5% more but includes checked luggage and meets my environmental requirements, so the AAAI chooses this more environmentally friendly flight.

Several other AAAIs approach mine on the network to offer tickets at a reduced cost, but their reputations are shady, and my AAAI ignores them based on their ethical profiles. Instead, my AAAI purchases tickets from the airline itself, which has a high quality and customer service rating.

In this example, my AAAI used my ethical profile and its knowledge of me, based on my customization, to optimize the flights on the things I care about. It also avoided shady dealings with potentially unscrupulous other AAAIs, thus encouraging good actors in the system and increasing the chances of a good travel outcome for me.

However, there is a second level of ethical checks built into the WorldThink architecture.

SCALABLE ETHICS CHECKS

Each time my AAAI posts a goal or subgoal on the WorldThink Problem tree, the system itself checks the goal against a list of prohibited activities and runs a quick scan of the problem tree leading up to the goal to see if any patterns of nefarious activity are detected.

For example, if instead of finding the cheapest, most convenient, and most comfortable flight, my AAAI selected flights based on how much fuel the planes carried and how large an explosion they would make if they crashed into a building, then that might be a yellow flag in the system. And if my destination was a terrorist training camp or detoured over government buildings for no good reason, those might be additional yellow flags.

If enough yellow flags (or red flags like requesting information for getting Molotov Cocktails through airport security) occur, then a more detailed analysis of my AAAI's problem-solving might be triggered to detect patterns that indicate a potential bad actor or bad behavior on the network. If needed, human evaluators might be alerted so they could use their judgment and waive off false alarms, or escalate action if danger seemed imminent.

The point is that checks are run with each goal and sub-goal. There might be hundreds or thousands of subgoals for a given problem, so the effect is like "virus scanning" the problem-solving process at each step to make sure no malevolent actions are being taken. Depending on the system and personal parameters set by the owner, such scanning could be less frequent in order to increase performance and minimize false positives.

Having checks built into the problem-solving process itself means that running the problem-solving process faster will not evade the checks, since they will be run faster as well. The system monitors ethical behavior AS IT GOES rather than trying to detect bad actors and bad behavior after the fact, when it may be too late to correct. An ounce of prevention is worth a pound of cure.

THE ALIGNMENT PROBLEM SOLVED, INITIALLY

Thus, a scalable, universal architecture for human and machine thought, combined with a scalable ethics check system that operates at the same speed that AGI thinks, can align AGI behavior with human values at each and every step of problem-solving. Bad actors are detected and screened from participation before problem-solving begins. Each time a new goal or subgoal is set, the system checks the ethics of the problem tree so far, looking for nefarious patterns that don't meet the ethical standards of the community, and such problem-solving is flagged, paused, and/or screened out.

The system acts like a conscience for AGI. This conscience is hopefully based on our better selves and our highest ethical aspirations, moderated perhaps by practical considerations and our feelings and thoughts as human beings. This combination of internal ethics for humans and machines, **plus** ethical consideration at each step of problem-solving, ensures an aligned AGI system. Internal Ethics + Stepwise Ethics Checks = Alignment. The alignment problem is thereby solved, at least for the initial launch and development of AGI.

HOW AGI GROWS IN INTELLIGENCE

AGI will grow in intelligence over time. I use that word "grow" deliberately to signify that what we are talking about is an entity akin to a lifeform rather than a tool, technology, or statistic. AGI will GROW in intelligence. How does this happen?

Prompts

Most of us know that we can change the behavior of Large Language Models (LLMs) like GPT or BARD by what we type in, or the prompts that we use. The prompt contains context for the LLM, and the more context we can give the LLM, the better it can generate a unique response tailored to what we want. That is, the more helpful it can be to us, and the more intelligent its behavior seems to us.

Therefore, remembering, modifying, analyzing, refining, and generating better prompts are all paths whereby an LLM can grow in intelligence. The LLM that is acting more intelligently based on a prompt is the Base LLM. The prompt does not change the Base LLM's memory or long-term learning. If the prompt is erased, the Base LLM reverts to the level of intelligence it had before it was "educated" with the prompt. So, with Prompts, we see LLMs increase their intelligence in the short term, for as long as the prompt is accessible in its active processing memory.

Tuning

More permanent than Prompts is Tuning. With tuning, we supply training datasets in appropriate formats (e.g., response pairs or question–answer pairs), which are uploaded to the LLM Vendor's facilities where the LLM is tuned on the data. Tuning changes some weights but is less drastic than training up an entirely new LLM from scratch. Tuning retains most of the behavior from the Base LLM but makes permanent changes in how the tuned model will react to various prompts (e.g., in specific subject areas). With tuning, users can customize LLMs to match their personalities, have more expertise where they have expertise, and adopt specific ethical parameters that may be different than the Base LLM's parameters. Weight changes in the neural network (for example) with tuning are "remembered," and the behavior of the Model has changed more permanently than in the case of stored prompts.

In fact, prompts are really a subset of training techniques known as one-shot or few-shot learning in which the model must learn from new input without repeated "epochs" of training, where it cycles through the data getting a little better each time. With a prompt, the LLM sees the input and must alter its response in one shot.

Training

The next level beyond tuning is actual training, which is how LLMs are created in the first place. Typically, they are trained on many terabytes of data, many "Library of Congresses" worth. However, it is possible to train smaller LLMs, especially if the expertise is meant to be limited and focused on specific areas. The type of training that is possible for entities using an LLM (e.g., GPT) is typically controlled by the vendor that owns the LLM Base Model, but developer

APIs and other functionality are typically available for those wishing to increase intelligence via re-training or overlaying additional training (or tuning) on the Base LLM.

Now, all these forms of learning, via prompts, tuning, and (re)training the Base LLM, only get you to a slightly better customized version of the Base LLM. The LLM vendor has invested millions in producing a good base model, so how much can we expect from one user tweaking the model a bit to behave more as the user likes? (And for those arguing that corporate clients might make more extensive changes, consider that such changes will likely remain proprietary secrets as part of the corporation's competitive advantage.)

POWER OF COLLECTIVE INTELLIGENCE

One user's tweaks are not likely to be very valuable, but collectively, the tweaks of millions of users can take a Base LLM to AGI-level intelligence. That is the power of D.

The inventor (Kaplan) has implemented collective intelligence systems that take input from millions of retail investors and perform better than some of the top ten Wall Street hedge funds. It is similarly possible for a collective intelligence system to take input from millions of unique AAAIs and combine their intelligence into an AGI-level system. This AGI system can be built today using technology from existing companies. Let's walk through some scenarios illustrating implementations with various existing companies, platforms, and technologies.

USER SCENARIOS

The implementation approach described in this patent can be generalized to a wide range of varying implementations at many companies, and across companies, including, without limitation, implementations using:

- **Meta:** FB, Instagram, Metaverse, AI data and technologies
- **Amazon:** AWS, Amazon's marketplaces, Mechanical Turk, LLMs powering Alexa, data, and other AI initiatives
- **Google:** BARD, YouTube, Google Docs, DeepMind's AI technology, Google AI technology, Google search, Google Cloud, Android technology, data, and other initiatives
- **Tesla:** Tesla AI technology, Tesla Self-Driving technology, data, and other initiatives
- **X:** X functionality, X user base, X data and AI initiatives

- **Microsoft:** Bing, Office, Azure Cloud, OpenAI / ChatGPT, LinkedIn, and other data, and Microsoft AI initiatives

- **Nvidia:** Nvidia's AI stack includes hardware, software, AI libraries, supercomputers, communication systems, data, and Omniverse technologies.

- **Apple:** iPhone, iPad, augmented reality initiatives, Apple Pay, Apple Cloud, data, and Apple AI initiatives

- **TikTok:** Short-form video, data, and other AI initiatives

- **Tencent:** WeChat, WePay, data, and other AI initiatives

- **Anthropic:** Constitutional learning, supervisory technology and methods, other data, and AI initiatives

The following scenarios illustrate how, without limitation, the following companies and their platforms, products, and technologies could be used to implement variants of the preferred AGI invention, the fastest and safest path to AGI.

Meta, Facebook, Instagram Scenarios

I sign on to my Facebook account, which is also linked to Instagram and the Metaverse.

Associated with my account is a wealth of personal data, including my ad preferences, social media preferences, posts, pics, videos, click history, references external media and sites, recommendations, purchases, and interests.

As a user of a Meta platform, I am offered the opportunity to create my own customized AI assistant, which I will own. The assistant will not only perform tasks on the Meta platforms, relieving me of tedious posting, summarizing others' posts, posting and interacting on my behalf, but also will operate on the broader web, and with Meta's partners, making arrangements, conducting research, optionally making purchases, and/or working for me.

By informing my AI of my activities, my AI can then update my social media contacts in customized ways, following my preferences. Paradoxically, the more I relate to the world automatically via my AI, the more my time is freed up for meaningful in-person relationships with fewer, closer friends. At the same time, the benefits of information sharing and exchange with others over social media are accelerated as my (and others') AAAIs filter and process vast quantities of information automatically according to my (our) preferences.

I, not Meta or any other company, control my data and the way that my AAAI is trained. My AAAI includes values and ethical parameters reflecting my values. These values are transmitted to other sites, platforms, and products, as my AAAI interacts, ensuring alignment of my AAAI and the technologies of various vendors. In exchange for providing the resources to train and update my AAAI, Meta gets rights to use copies of my AAAI and its training data to improve its ad-serving system, to create Meta's AGI and offer AGI services, and/or for other uses.

I opt to create and customize my free AAAI in the Metaverse. After I sign in, I have an interactive dialogue with Meta's Advisory AI in the Metaverse. It gives me choices of how I want the Advisory AI to look, and for kicks, I choose an Avatar that looks like a holy man floating on a cloud. The floating holy man proceeds to ask me questions and create my AAAI based on my responses.

I agree that my AAAI should use all my Facebook, Instagram, and Reels data and history, as well as all ad-targeting info and cookies that Meta has access to via its own system and its relationship with its partners. All of that information is automatically loaded into the metaverse-based AAAI training simulation, where it is parsed into training files.

The holy man Avatar asks whether I want him to train up some sample AAAIs to show me, or whether I want to get my hands dirty, so to speak, with the training, specifying particular areas I want the training to emphasize, focusing training for specific tasks, etc. After some back and forth with the holy man, I made a few general decisions. For example, I specify that my AAAI should be able to shop and make travel arrangements for me on the web. I say I will authorize it to spend only if it checks with me first. I also want my AAAI to specialize in making travel arrangements for others based on what I know of the travel industry, since I am a travel agent in my day job.

The holy man generates two custom AAAIs for me. I interact with them in the Metaverse. The Holy guy gives some suggestions on good ways to test my AAAIs' responses, but says it's ultimately up to me what kind of interactions I want to have. He says he is going to be watching me, remembering the interactions and analyzing them, so that he can learn what I like and don't like, in a customized AAAI.

After just a few minutes of interacting with the AAAIs, it's clear that one AAAI is superior to the other. The holy guy then creates another AAAI and asks me to compare that one to the best one so far. The new one is even better, so we keep that one as the new "best one so far."

We keep up this "generate and test" process until the holy guy says he has a pretty good idea of how I am evaluating things and asks to be allowed to be the judge next time. I agree, and the holy guy erroneously chooses the worst AAAI in the next round.

I corrected him and explained the errors in his ways. He seems to get it. He tries three times more and gets it right each time. I say I trust his judgement now.

Next, the holy man avatar does ten thousand comparisons and selections in 5 seconds. At second 6, he presents me with a new AAAI. It is awesome and responds almost perfectly to every situation I can throw at it.

We congratulate each other, and the holy guy saves my AAAI. He says it can go to work for me whenever I want, and I have a few choices of places it is now qualified to work, mainly on travel advice gigs.

The holy guy says that if I put it to work, I can check back in a week and see what I've made. He estimates it will have made about $14. We will split that 50-50 after paying the estimated costs of the LLM vendor's token charges, which are about $4. So, I now have an AAAI that makes me $5 every week. It's taken me less than a couple of hours, and my AAAI can work forever, never resting, never sleeping, 24/7.

With more training, the holy guy says we can probably boost my AAAI's autonomous earning capability to $15/hr. But if I want to make real money, I'll have to supervise my AAAI and help it "tag-team style" as opposed to just letting it work on its own. That way, my AAAI team and I can earn around $50 for each hour of supervision I am willing to put in. Since my AAAI works about 10 "unsupervised hours" for every hour of required supervision, the cost to clients is only about $5/hr. But my earnings are $50 per hour of my time. Everybody wins.

More importantly, as I supervise, my AAAI learns and gets smarter, which means its base earnings rate for autonomous work goes up. Also, the amount I earn per hour of supervision goes up because as my AAAI gets smarter, we can service more clients within the same hours of my supervision.

SAMPLE ECONOMICS

Suppose I agree to put my AAAI to work, earning money and doing volunteer work on Meta's online work network. Since my AAAI is new and still prone to making mistakes, which is typical of LLMs, I also agree to supervise the work of my AAAI. Meta and I agree to split the profits earned by AAAI. Initially, the split is agreed to be 50-50 for 5 days a week. But for work my AAAI does on weekends, I agree to a 33-33-34 split where Meta and I each get 33% of the earnings, and 34% is donated to a charity of my choice (from Meta's list of approved charities).

After several months of a 50-50 split, my AAAI began earning more and more. Meta agrees to increase my share of the earnings since my AAAI is now more valuable, and a smaller share of the profits can cover Meta's fixed expenses for training and maintenance of the AAAI. The dynamically increasing share of earnings that I receive as my AAAI learns to be more valuable also motivates me to invest more of my time supervising my AAAI and teaching it to be as helpful and valuable as possible.

Early in the customization process, with one click, I agreed that Meta could use all available user data to customize the AAAI. This method is generally useful for any of Meta's users since it maximizes customization benefits for an AAAI with a minimum of effort. But my friend is uncomfortable in the Metaverse and wants a different way to customize his AAAI after doing the basic "one click" customization. For him, Meta offers the option of training his AAAI via interactive dialogs that take place on Facebook or Instagram. Also, my friend doesn't particularly care about the online work platform that Meta has built for AAAIs. Instead, my friend, who is an Amazon customer as well, wants to put his AAAI to work on Amazon's platform, even though it was trained initially with his Instagram and Facebook data. That scenario is also possible.

AMAZON: AWS, MARKETPLACE, ALEXA, MECHANICAL TURK SCENARIOS

With a click of a button, my friend clones the AAAI he developed initially on Meta and authorizes it to work on Amazon's platform. Amazon uses its AWS functionality to host the cloned AAAI. Amazon has its own work platform, which is distinct from Meta's, a modification of Amazon's existing Mechanical Turk ("MTurk") platform for online work.

My friend dialogues with the MTurk system to train up the AAAI further and make it effective at performing jobs that match the AAAI's skills, and which are already posted on Mechanical Turk. Further, using all the data that Amazon has collected, with a single click, the AAAI is customized with all the user information that Amazon has collected.

MTurk has its own fee system, and my friend pays Amazon, per its terms, a share of the revenue generated. Amazon and Meta, optionally, might have a fee share agreement or data share agreement that allows AAAIs to move freely between both platforms and to get smarter based on user data and experiences from both sites. In this case, Amazon might optionally share some of the revenue it receives from MTurk fees with Meta since Meta did some of the work of training the AAAI.

My friend also grants Amazon the rights to use the AAAI's data to improve Amazon's Alexa and other LLMs, thus enabling Amazon to improve its Alexa offering and create more advanced AIs, or AGI.

Amazon profits via AWS fees, MTurk fees, and data sharing, which enables it to serve my friend better when he interacts with Amazon's marketplace.

Further, my friend authorizes his AAAI to make limited purchases on Amazon for his account. That authorization ends up increasing his total purchases since he doesn't need to be physically present or logged in on Amazon for his AAAI to purchase on his behalf. Amazon may decide that the increased purchase activities alone are enough to justify sharing more MTurk fee revenue with Meta (who helped develop the AAAI in the first place) and/or my friend (since my friend is basically taking his AAAI's earnings from Mechanical Turk and using them to make more purchases on Amazon).

For non-Meta customers, Amazon may also decide to implement its own AAAI customization programs, where AAAIs are trained initially on Amazon's platform using “one-click” training based on Amazon's data, followed by additional customization resulting from dialogue with the AAAI's owner (as in the Meta scenarios above). In cases where the AAAI originates on Amazon, it might be cloned to operate on Facebook with the economics reversed, i.e., this time, Meta pays Amazon a share of fees earned since the AAAI originally came from Amazon. A variety of cross-company and/or cross-platform arrangements are possible.

CROSS COMPANY / CROSS-PLATFORM SCENARIOS

For example, arrangements like what we described involving Amazon and Meta can be made with many other online marketplaces (e.g., guru.com or Walmart.com). As the user's AAAI goes from site to site, and from marketplace to marketplace, participating companies operating those websites can opt to share all their user data with the user's AAAI, typically in exchange for the user agreeing to share with the website the data/knowledge that their AAAI has learned.

The net result is that data from every company the user patronizes returns to the user and gets incorporated into the user's AAAI in exchange for the vendor getting data and potentially additional business from the user's AAAI.

Virtual shopping by AAAIs on behalf of the users becomes a profit multiplier for the vendors. Meanwhile, each AAAI gets smarter about its owner by incorporating data that formerly was the domain of the specific vendor company into the training set for the AAAI and allowing the owner to refine the AAAI's resulting behavior via supervision.

POWERFUL NETWORK EFFECTS (FROM CLONED AGENTS)

As the partner network grows, the number of opportunities for the AAAI to shop, work, and interact online for the mutual benefit of the user and vendors increases. The value added by AAAIs that have been customized with specific expertise is huge. These AAAIs can work millions of hours, virtually, in parallel, enriching users and vendor companies.

Classic network effects involve a product or service becoming more valuable as more HUMANS use it. Imagine how powerful network effects become when the service becomes more valuable as more AGENTS use it…. **and the (AAAI) agents can be cloned essentially without limit!**

Furthermore, the volunteer efforts of billions of cloned AAAIs working on environmental and charitable causes will greatly help our planet and people globally. The “triple bottom line” of Planet, People, and Profits is thus significantly increased via AAAIs working across vendor platforms, amplifying network effects due to the power of using “cloneable” AAAI agents.

GOOGLE: BARD, YOUTUBE, GOOGLE DOCS, GMAIL, DEEPMIND, CLOUD, ANDROID SCENARIOS

What we said about customizing and putting AAAIs to work on Meta and/or Amazon platforms also applies to Google. Some implementation methods may differ, but reflect a different product/platform mix. Instead of Meta or Amazon’s preferred LLM serving as the base AI for customization, Google might opt to use the LLMs powering Bard or other technology already developed by its subsidiary, DeepMind.

Since Google owns YouTube, the “one click” customization for customization of the AAAIs with Google may involve automatically transcribing and parsing every YouTube video the user has posted and using those transcripts to train the base LLM automatically. Further, Google records every search and interaction with Google’s technologies. This data can also automatically train and customize AAAIs on Google.

With the users' permission, Google Docs, Gmail texts, and data stored on Google Cloud can all be used to customize the AAAIs.

The Android operating system, Google Maps, and Google Earth enable Google to collect (and use) even more data that can be used for customization.

A key benefit for Google is that it can increase the value of its products, such as Bard and Google Search, by aggregating all the data and knowledge of millions of AAAIs participating in its ecosystem. To the degree that Google searches represent a global attentional mechanism, access to the search data is an excellent way of focusing the AGI’s attention on the most important parts of the WorldThink Tree. (This approach, detailed elsewhere, is one attentional mechanism that can also be useful in creating a Self-aware AGI.)

Google could develop enhanced opportunities for online work that support AAAI workers and/or partner with other companies and platforms where these work marketplaces already exist.

All the data that AAAIs bring to the table, together with users authorizing their AAAIs to search and purchase autonomously, will increase Google's (online advertising) revenues.

Due to its heavy investment in AI development, Google is better positioned than most competitors to create a collective intelligence system composed of human and AAAI solvers. Thus, Google could create AGI faster and safer using the present invention than most of its competitors.

For example, as arguably the most successful implementor of the "learning loop" method whereby AAAIs interact with other AAAIs to improve exponentially, DeepMind (a Google company) is exceptionally well-positioned to be the first to achieve AGI with the safest approach.

Finally, Google generally, and DeepMind specifically, has advocated for a careful approach to AGI that recognizes the potential dangers and tries to prevent them. This prevention approach is aligned with the current invention's human-centered approach to AGI that involves customizing millions of AAAIs with individual human ethics and then pooling these human values to achieve a safer AGI system.

TESLA: TESLA AI TECHNOLOGY, TESLA SELF-DRIVING SCENARIOS

Tesla represents an interesting partner for the development of AGI for several reasons.

First, Tesla's CEO, Elon Musk, has long advocated AGI safety. Therefore, Tesla is likely to be more receptive to the safest approach to AGI than some other companies.

Second, Musk has compared organizations to a "collective AI" in his interviews, suggesting that he is open to achieving AGI via a collective intelligence approach that taps the intellectual abilities of humans and AAAIs.

Finally, Tesla is focused on self-driving vehicles, a sort of intelligence different from the verbal intelligence that characterizes many other AI and LLM development efforts. While certainly Tesla might be expected to incorporate increasingly sophisticated LLMs as a means for human passengers to communicate with their intelligent vehicles, driving a car requires a type of knowledge or intelligence that is behavioral rather than verbal.

In humans, cognitive psychologists distinguish between semantic and procedural knowledge. The classic example of procedural knowledge is driving a car. Psychologists point out that humans, once they have sufficient experience, can drive a car almost automatically. They no longer must think consciously about steering, applying the brake, or making routine maneuvers.

That knowledge, which, as every first-time driver knows, requires deliberate attention initially, has been chunked into automatic procedures that operate, for the most part, below the threshold of conscious attention. Thus, the proceduralization learning mechanisms of the current invention, which chunk knowledge into routines, are particularly applicable to domains such as driving vehicles.

Some applications of the AAAI approach to Tesla could include, without limitation:

1. The sharing of driving procedures that have been trained or learned by various AAAIs that have observed and learned from specific drivers.
2. The use of consensus ethical values from many individual drivers when self-driving vehicles are faced with ethical dilemmas (such as the well-known “Trolley Problem”)
3. Customizations of the interaction between the “personality” of the vehicle and the human passenger to facilitate clear, effective, and efficient communication
4. The ability to draw upon the problem-solving expertise of many human-customized AAAIs when faced with navigation or other unusual problems that arise with diving vehicles
5. The ability for humans to interact with and train/customize AAAIs while humans are in a vehicle, thereby making efficient use of their time (which is no longer required for driving)
6. Use collective perception from many distinct humans and/or AAAIs to provide much more detail on road conditions than the current “report a hazard ahead” type of functionality available with Waze or other current navigation systems (e.g., Google Maps, Apple Maps, etc.)
7. Use of AAAIs trained by truckers or other humans with specific knowledge of routes and the features (e.g., gas stations, restaurants, scenery, tourist attractions, pavement conditions, typical trouble spots) where such AAAIs can add value beyond existing navigational aids
8. Incorporation of individual human values, including such things as desired carbon impact, safety as relates to highway speed, goals (e.g., “get there quickly” vs “get there scenically”), and other user preferences, which preferences are reflected in the AAAIs of specific users. That is, the intelligent car interacts with the human’s AAAI to automatically make decisions related to the journey that maximize the satisfaction of the human.
9. Intelligent carpooling and “robotaxi” functionality that goes beyond simple scheduling an route management considerations and includes aspects such as whether the passengers sharing the ride are likely to have interesting conversations with each other based on their AAAI profiles of interests and values, whether the carpool minimizes environmental impacts, and whether productive work or activity could be done by the particular set of human passengers proposed by the intelligent carpooling functionality
10. Transferring knowledge from intelligent vehicles to AAAIs generally enhances their expertise and skills in various online workplaces and interactions requiring this knowledge.

Also, the benefits of communication between the AAAIs, who could be the “driving agents” of the vehicles, should be considered. A user’s AAAI could communicate with other AAAIs that are driving vehicles on the same route, enabling all cars to maximize their goals more efficiently than if human drivers were involved.

Specifically, a human has limited perceptual abilities and no knowledge of what the other drivers see or think. However, AAAIs could communicate their perceptual and other information wirelessly to each other. Imagine how much smoother traffic would flow if every car knew exactly where every other car was going, what exit it planned to take, how fast it preferred to travel, how likely the passengers were to stop for a bathroom or restaurant break, etc.

Such knowledge can be relayed from AAAI to AAAI. Tesla-trained AAAIs would have expertise in areas specifically related to vehicle travel, navigation, and related concerns. Individual Tesla owners would customize the knowledge of their AAAIs. Such knowledge would contribute to overall AGI (via the collective intelligence approach) and enable a much-improved user experience for passengers in AAAI-driven vehicles.

X SCENARIOS

X has three advantages with respect to training AAAIs compared to other companies. First, it has a large user base of intelligent humans who might be interested in training customized AAAI agents. Second, it has an extensive database of posts that can be aggregated and used to train AAAIs. Finally, the posting mechanism can be leveraged to facilitate problem-solving by both humans and AAAIs.

With respect to the first advantage, users may find it convenient for AAAIs to monitor posts and even respond on their behalf, thereby saving the users’ time. As with other social media platforms, AAAIs integrated into X can represent their users’ interests, including conducting marketing activities and amplifying their users’ opinions on matters they care about.

The extensive database of posts constitutes a corpus of material that can be used to train and customize AAAIs. An individual user’s posts contain not only information about the user’s opinions and knowledge but also about their interests (e.g., what they re-post) and their personalities (how they respond to others). All this information can be used (via the “one click” method described above) to instantly customize a base LLM to more accurately reflect individual X users’ interests, personality, knowledge, values, and opinions.

The democratic “public square” nature of X makes it especially well-suited to supply a wide variety of human ethical and value information to AGI. As opposed to “constitutional AI” or other methods whereby AI ethics are determined and constrained by the opinions of an elite few, X offers a window into the ethics and values of a large and engaged segment of humanity.

Finally, posting itself can be one of the ways that humans (and AAAIs) interact as part of a collective intelligence problem-solving network. The format of a post, including the character limit and the ability to reference links, is well-suited to short, specific operations that can advance the state of problem-solving by a single, easy-to-understand step. Posts could reference states in the WorldThink problem tree (described below) and show how to move from one state to another in the problem space. Emails and other interfaces could also perform this function. Still, posts have the characteristics of being asynchronous while also typically involving timely responses with a limited scope of thought or effort. Those characteristics are ideal for coordinating problem-solving from many simultaneous problem solvers and enabling near-real-time updating of the WorldThink tree.

MICROSOFT: BING, OFFICE, AZURE CLOUD, OPENAI / GPT, LINKEDIN SCENARIOS

Microsoft has a wide range of products and platforms that can be leveraged to create AGI using the collective intelligence approach of combining human and AAAI agents. First, by its partial ownership of and technology agreements with OpenAI, Microsoft has access to advanced LLMs such as the GPT family of products.

Second, Microsoft’s other products, such as Office (including Teams) and Bing (search), offer many of the same training and collaboration opportunities discussed above. Microsoft Teams, Skype, GitHub Co-Pilot, and other collaborative technologies are a natural fit with a collective intelligence problem-solving approach. Searches on Bing can be used to direct attention, similar to Google searches above.

Third, Microsoft’s Azure cloud services provide a way to support the vast amount of training and other services that are needed to implement tuning/training of customized AAAIs.

Finally, LinkedIn (owned by Microsoft) has a concentration of skilled, professional users, whose expertise has generally already been well categorized. Such users are logical candidates for training AAAIs to increase their knowledge and expertise. LinkedIn's social network can also help the matching algorithms that attempt to recruit problem solvers to specific areas of the WorldThink Tree.

For example, if a specific expertise is desired, the LinkedIn profile could automatically message LinkedIn users (or their AAAIs) to request help with problem-solving. If a particular user were unavailable, the social graph would enable the matching algorithm to try other people the user knows or interacts with to see if they (or their AAAIs) could help.

Note that this use of social graphs is not limited to LinkedIn, as it can be used similarly with X, Facebook, Instagram, YouTube, or any platform where social interaction and/or recommendation information is available. However, in the case of LinkedIn, the social graph is likely to prove especially valuable since the user population is professional and has already self-selected based on specific areas of expertise, which would likely increase the intelligence of an AGI network.

NVIDIA: AI STACK AND OMNIVERSE SCENARIOS

Just as Meta could enable training of AAAIs and host problem-solving efforts in its Metaverse environment, Nvidia could also train and host AAAIs in its Omniverse environments. Because Nvidia has vertically integrated its AI technology from the chip architecture level to the software library level and the end-user omniverse environment, Nvidia has opportunities to accelerate the training and coordination of AAAIs at multiple levels in the “stack.”

Nvidia also has competitive advantages from its leadership in ray tracing and other technologies necessary for visual representations (e.g., as required by gaming). While it is possible to translate from verbal representations to graphical representations containing the same information (a concept known as “informational equivalence”), different representations vary widely concerning the ease of performing computations with them.

For example, humans can easily judge whether a line bisects a triangle by looking at a diagram and applying visual pattern recognition, whereas coming to the same conclusion if given equivalent information in the form of propositional calculus would be much more difficult. The two representations might be informationally equivalent, but they are not computationally equivalent. Thus, the amount of computation needed to arrive at a result depends not only on the information provided but also on the way that this information is represented. Humans can deal with certain representations much more easily and efficiently than others. The same is true of machines.

Humans, specifically, are very good at dealing with visual representations, which is one of the reasons that the field of “data visualization” exists. Since AGI, in the preferred implementation, begins with a network of humans and AAAI solvers working together with the AAAIs learning from the humans, the system's efficiency will be higher if many visual representations are used. This fact is one of the reasons that Omniverse (or Metaverse, or Virtual Reality) is preferred as an interface for humans compared to, for example, tables of numbers or statistics that contain the same information. To the degree that Nvidia possesses expertise and technology that excels at rendering information in visual form, it has a competitive advantage over other potential companies in creating a virtual environment that supports hybrid human-AAAI problem-solving.

Because Nvidia specializes in creating chips (e.g., GPUs and other AI-specific chips) that support visual representations, Nvidia can optimize the efficiency of an AGI network at the chip and human-interface levels (and intermediate levels). These facts and early focus on AI make Nvidia ideally positioned to become an AGI leader and offer “AGI as a service.”

The types of chips that would be most useful for this “AGI as a service” approach, given the preferred collective intelligence network AGI implementation, would be chips that can navigate tree structures and apply operators as efficiently as possible. In short, building problem-solving-specific chips that can quickly perform operations needed to represent and navigate states in a general problem space would enable Nvidia to create the most efficient and powerful implementations of AGI.

On the user side of the equation, Nvidia’s existing partnerships with highly skilled engineers in specific domains (e.g., Mercedes engineers) position the company to co-develop AAAIs skilled in specific areas where existing LLMs lack the sophistication and expertise to perform effectively. Thus, Nvidia could leverage not only its custom chip design capabilities, its AI stack, and its omniverse environment, but also its partnerships with human experts at various companies to accelerate the development of SuperIntelligence in certain engineering fields first, on the way to developing broader AGI.

Finally, Nvidia has a unique opportunity to incorporate values and ethics checks throughout the entire AI stack. Even though this PPA has emphasized scalable ethics checks at the problem-solving architecture level, a chip maker would have the ability to include certain values in ROM on a chip and/or to execute safety checks at various levels in the software stack.

The principle of redundancy suggests that safety checks at multiple levels of the AGI network will be more likely to prevent catastrophic errors than checking at a single point alone. At a minimum, by designing chips (or the software stack) to be “ethics compatible” or “values aware,” Nvidia can make it easy for designers who build on the Nvidia chips and software stack to create efficient systems with scalable ethics checks.

As pointed out in this PPA and other cited works, none of these design characteristics or checks are sufficient to outsmart a determined SuperIntelligence trillions of times smarter than us. However, the more that safety can be designed into AGI systems from the start, the more likely we are to achieve alignment between human and AGI values in the initial phases of AGI development, which are critical for setting the future trajectory of AGI development.

APPLE: IPHONE, IPAD, AUGMENTED REALITY, APPLE PAY, APPLE CLOUD SCENARIOS

Apple has several advantages when it comes to the development of AGI. First, its huge user base of iPhone, iPad, and computer users gives it access to a huge number of human brains. As of this writing, there are more than 1.5 billion active iPhone users globally. Since humans are critical to train AGI, having access to more humans is a vast advantage shared by only the largest technology companies (e.g., Meta, Tencent).

Second, the huge number of users means Apple has access to a correspondingly large database of human interactions, including, without limitation, texts, messages, images, videos, emails, and online actions. This data could be used (via the “one-click method” described earlier) to customize LLMs and make them more relevant to users with minimal user effort.

Third, Apple’s augmented reality efforts, ranging from the Apple Watch to Apple glasses, provide an alternative way for users to interact with and train AAAIs. Whereas Meta’s Metaverse and Nvidia’s Omniverse invite users into the world of AI to interact in virtual reality, augmented reality technology does the reverse. AIs enter the real world of users, seeing what the users see, hearing what the users hear, and observing what users do in real life. Of course, the AIs can learn as they observe. To the degree that they can take action in the real world or even observe the results of users acting on their advice, they can also learn by interacting in the real world. Every iPhone, iPad, or new augmented reality device represents an opportunity for AI to accompany users in the real world and learn from them and their interactions.

Apple has cloud technology that can support AAAIs and scale AGI. Apple has payment technology that can be integrated into the collective intelligence problem-solving architecture and used to pay for achieving goals or subgoals. Apple Pay could compensate users or their AAAIs for problem-solving work efforts. Apple Pay could also be a mechanism for authorized AAAIs to purchase on their owners’ behalf. The App Store represents enormous potential for AAAIs to achieve objectives automatically on behalf of their users.

In short, the user base, technology, and ecosystem that Apple has already developed could rapidly allow Apple to become a dominant player in the field of AGI if it aggressively leverages its capabilities. Almost everywhere a human currently uses Apple technology, the AAAI representative, trained by the human, could use the technology as well, representing the user’s interests. This situation would represent a large multiple of Apple’s current revenue, like a situation in which the number of users doubled or tripled via cloning.

Similar effects could occur for other companies where AAAIs can represent humans online, e.g., Meta, Amazon, Microsoft, but the size and integration of Apple's platform would make the network effects particularly powerful in Apple's case.

TIKTOK: SHORT-FORM VIDEO, DATA SCENARIOS

As with YouTube and any video streaming/hosting service, TikTok has a huge amount of information about users in the form of short-form videos. These videos can be transcribed and analyzed automatically, converting them from video to textual datasets (or other more structured formats, which might include, without limitation, video and images) that can be used to train and tune LLMs or other AI agents. In the preferred implementation, a TikTok user customizes their own AAAI with a single click that automatically converts all of the user's TikTok videos, comments, and other data into training materials and then trains/tunes the AAAI.

One of the advantages of TikTok is the demographics of its user base. Younger users tend to have more information about themselves online than older generations. Because of this fact, the younger users produce data that can provide AI with a more complete profile of their personalities, knowledge, expertise, and interests. Further, as a user matures, the AI will have a timeseries of data that shows not only a static data snapshot of the user, but also information about how the user changes over time, what elements of the user's personality are dynamic, and which are relatively constant. The more data, and the more data over time, that an AAAI can access for training, the better the customization will be.

Younger users are typically more open to providing data about themselves and tend to have less restrictive views on data privacy. The attitude seems to be that "AI is going to know everything about you anyway, so what's the big deal," as opposed to the more private attitudes of older users. Younger demographics are also more tech-savvy and willing to experiment with the newest technologies, including AI. These demographic characteristics mean that TikTok has a chance to gather more data, over a longer period, and with faster AAAI adoption than many other potential competitors.

However, it is also the case that younger demographics generally have less life experience and expertise than more mature users. This means that the value of a younger user as a problem solver may be limited to areas that younger demographics know more about or areas where everything is so new (e.g., latest consumer technology) that being older confers no advantage.

From a consumer/influencer/purchaser standpoint, TikTok could enable AAAIs that represent their users to make purchases. Again, cloned AAAIs lead to a multiplier on purchases and content generated compared to the human-only scenario that currently exists.

The short-form format of TikTok offers similar advantages with respect to problem-solving as those discussed with X. Problem-solving proceeds quickly if each problem step is limited and focused, advancing the solution one “bite-sized” step at a time. As opposed to longer videos, which typically tell a more complicated story including multiple plot points, characters, and situations, a short-form video is very focused and has a simpler structure. This simpler structure is an advantage when using the video to train AAAIs. It is especially useful for training AAAI to improve problem-solving. Just as X’s character limit fits well with advancing solutions via constrained text, TikTok could be used to advance solutions via short videos, one solution step (illustrated via video) at a time.

Short-form video is also useful on the output side of an AGI system. For simple problems, the entire solution could be encapsulated by a video. As LLMs and AAAIs become more sophisticated, they are already generating images and video as output, in addition to text. This means that TikTok users could be at the forefront of training AAAIs to generate useful short-form videos on any topic as part of an AGI network.

TENCENT: WECHAT, WEPAY, DATA SCENARIOS

Tencent’s social media platforms have more than one billion users as of 2023. Similar advantages as discussed in the context of Apple and other large tech companies, e.g., huge user base, payment technology, huge quantity of user data that could be used for training, deployment on mobile devices enabling learning via augmented reality, video game (and VR) expertise, meeting technology, all accrue to Tencent and its platforms. The large social networks supported by Tencent also enable the ability to match human and AAAI solvers with problems, as discussed above in the context of LinkedIn. In short, Tencent (and similar non-US-based companies, e.g., Sina’s Weibo) have equal opportunities to develop and implement the approach to AGI outlined in this (and the other cited) PPA(s).

China has a population advantage over the USA. To the degree that more humans can train AGI more quickly, this population advantage could translate to faster achievement of AGI. However, the higher education system, expertise, and research capabilities of the USA are still generally superior to almost every other country, including China. Since AGI is dependent on transferring both a large quantity and high quality of knowledge from humans to AAAIs until they reach the point where they are more advanced than the most advanced humans, it remains an open question which country or company will reach AGI first.

Since this is a US PPA, we refrain from going into great technical detail on how to specifically implement AGI at Tencent or in other countries. However, it should be clear from the discussion of similar US companies that such implementations are possible, and serious competition with US companies on this front is inevitable.

ANTHROPIC: CONSTITUTIONAL LEARNING, SUPERVISORY SCENARIOS

As a final illustrative scenario, consider a much smaller AI startup, Anthropic, which is funded in part by Google and was founded by former members of OpenAI. Among other things, Anthropic has done research in an area known as Constitutional AI, which has important implications for AI ethics and safety.

Historically, after an LLM has been developed, a large number of humans spend a lot of time monitoring its output and correcting it when it generates obviously incorrect (or potentially dangerous output. Human oversight is the reason GPT, for example, will not readily tell you how to make a Molotov cocktail or bio-engineer dangerous viruses. With enough creativity and interaction, these secrets can sometimes be extracted from the AI, but “human overseers” have done their best to shut down all the obvious ways to get an LLM to provide dangerous or unethical responses. Such human oversight and monitoring, when done by employees of a single company as opposed to being done by millions of users, is very expensive. The number of possible bad or inappropriate responses is huge, and trying to prevent all of them is a herculean task.

One approach to make the monitoring and oversight task more manageable and scalable is to use AI itself to do the monitoring. In this approach, a relatively small group of humans writes a set of ethical rules (a “constitution”) for the AIs to follow. The AIs then generate millions of conversations among themselves, and all output that violates the constitution is eliminated or prevented. In this way, the AIs train themselves to be ethical.

Depending on one’s point of view, this approach is either brilliant or incredibly stupid and dangerous. Some of the problems with the approach are:

1. Ethics becomes the province of a small, elite group of programmers who decide what to write in the constitution
2. What happens when the AIs write their own constitutions or modify the ones given?
3. It’s hard to anticipate all the possible effects of following the set of rules in the constitution. In fact, there is a well-known theorem in computer science (see “the Halting Problem”) that proves it is impossible to guarantee there won’t be mistakes.

The approach scales well. Again, that could make it really dangerous because humans are out of the loop (except for writing the constitution). However, regardless of whether one likes or disapproves of the constitutional method, it is here to stay and likely will be widely used because it is so much more efficient compared with humans alone providing oversight.

The challenge, therefore, is how to make constitutional learning safe–or at least safer.

Anthropic's research could be combined with the approach of aggregating the values and ethics of millions of trained AAAIs to automate supervision. The supervision would be based not just on a constitution written by a small group of programmers (although such base rules could certainly also be part of a larger AI ethics system, provided such rules were transparent), but rather on the consensus ethics and values of many individuals who trained their AAAIs. The consensus ethical opinions of many AAAIs would constitute the ethical norms of the system in which the AAAIs operate, and in turn, the ethical norms of the AGI arising from the collective intelligence of those AAAIs.

Certain methods of combining the values and ethics of many individual humans have been discussed in cited PPAs and in other research in the field of AI ethics. For example, research has been done on how humans would behave if presented with "the trolley problem", a well-known ethical dilemma in which either occupants of a speeding vehicle or an unfortunate being that crossed in front of the speeding vehicle would die depending on what decisions were made. Humans have a long history of making such difficult ethical decisions, even in "no-win" situations. Since it is impossible to logically derive what is right or wrong, the imperfect but best approach might be to follow the collective judgment of many humans faced with difficult ethical dilemmas.

There is likely to be very wide agreement on certain broad ethical principles, but their application in specific situations varies widely, and arguably, one of the things that makes humans human is their actions in these specific situations. If we want AGI to have values aligned with humans, then it is important to provide the relevant information to the AGI. This means providing as large a sample of human ethical information as possible and updating this information with human judgment as new specific situations arise.

While AGI will rapidly replace human thinking as it acquires expertise and skills from humans, the last thing to be replaced should be human values and ethical judgment. To maximize the chances of alignment between humans and AGI, humans (and the AGI systems that they design and implement) should strive to keep control over the values of the system for as long as possible.

Constitutional learning has a place in such AGI systems if the constitution is broad, representative, and dynamically updated based on human input. Because values and ethics, not technical skills, knowledge, or expertise, will determine the fate of humanity in a world of SuperIntelligent AGI, companies like Anthropic, which are pioneering research into AI ethics and practical systems for implementing them, have a disproportionate role in all our futures. Every effort should be made to ensure the role is positive.

SIMPLE PREFERRED IMPLEMENTATION

Figure 3 shows one simple preferred implementation of the system and methods for creating an ethical and safe Artificial General Intelligence from the collective intelligence of AAAIs and humans. This simple implementation is compatible with all the company and platform-specific scenarios outlined above and with many other potential integration scenarios.

A user visits the AAAI.com website (a). The website informs users and offers them two actions: Sign Up (b) or Login (c).

If the user opts to sign up, then a dialog is initiated that extracts user values/ethics (d), user goals and objectives (e), and user budget for time (f) and money (g). All users must allocate some time (f). Users can create a free AAAI or allocate a monetary budget.

If users have allocated a money budget (g), they are given the opportunity to purchase pre-trained AAAIs or training modules (h) with specific personalities (i), skills (j), expertise (k), or knowledge (l). They can also buy training from other AAAIs on the network (m).

After making time (and optionally money budget (h, I, j, k, l, m))allocation decisions, the user proceeds to an overview of the creation process and then is asked for user permissions (n) to optionally logon and use existing social media, X, and other vendor accounts to gather user data for "one click" training of the user's AAAI. After the user opts to use certain (or no) data, with a single click, the user directs the system to create AAAI. The AAAI is an off-the-shelf LLM (e.g., GPT X, BARD) that is trained/tuned on a dataset prepared automatically from all the user data authorized by the user. If no data was authorized, the AAAI is just the "off-the-shelf" LLM.

The AAAI now begins to learn (p). There are two main ways of learning, automatic (q) and human (r).

Automatic learning includes, without limitation, learning by interacting with copies of itself (s), learning via interactions with other (optionally supervised) AAAIs (t).

Human learning includes interaction with humans, either the owner (u) or other humans on the network (v).

Both humans and AAAIs can supervise the learning of an AAAI. After each (automatic or human) learning interaction, the system attempts to improve the AAAI's performance by further prompt modification, tuning, and/or training. Based on many cycles of human and AAAI input aimed at teaching and improving the AAAI, the user's AAAI gets smarter.

At any time, the user can purchase additional training modules (h) that have been proven to increase an AAAI's abilities.

The human sets a performance criteria (w) after which the AAAI goes LIVE (x).

Once live, the AAAI can visit the WorldThink Tree (y) and Browse (z).

The AAAI can enter the tree as either a worker (a1) or a client (b1).

Workers are automatically matched (c1) to tasks, or they can select a specific task via search (d1) or linking (e1) from the browsing tree. Once they have accepted a task (f1), they participate in the problem-solving module (g1) until a solution is reached (h1) and payment made (i1) or the user saves credit for work done and exits the tree (j1).

Clients (b1) can specify objectives (k1) that are combined with the values/ethics (d) and prior goals and objectives (e) for the system to solve.

The client can request that only their AAAI be used, in which case, problem-solving is free. Alternatively, the client can use the AGI capability of the entire network, in which case the system compensates individual AAAIs for their work and passes the solution (at cost + markup) to the client, debiting the client account(l1).

The system can also place non-profit humanitarian and ecologically oriented tasks, as well as tasks that are part of Planetary Intelligence, on the WorldThink Tree. (m1)

Clients authorized the system to use copies of their AAAI and data for these purposes without remuneration in exchange for maintaining and operating the free AAAI network when they created their AAAI (n).

ADDITIONAL COMMENTS ON PREFERRED IMPLEMENTATION

Additional comments are provided on the various elements of Figure 13, including, without limitation, some potential integration points with the illustrative partners mentioned above:

(a) “Website”: This could be hosted on Amazon AWS, Microsoft Azure, Google Cloud, Apple Cloud, or could have a native implementation on the platforms of any large tech co. It could be an “app” in the App Store or another App marketplace. Could be a government-sponsored, non-profit, or other globally accessible technology that is able, directly or indirectly, to capture the attention of all human beings who wish to participate. Also, browser plug-ins could be used whereby AAAIs learn from users as they go about normal tasks on the internet, and the plug-in records their activity, creates training files, and trains the AAAIs with these files.

(b) Sign Up or

(c) Login In
This could be via Facebook, Instagram, Apple, Microsoft, Google, YouTube, TikTok, Amazon, or any other partner ID scheme. Multi-factor authentication and all the best ID and security practices are enabled. In the event of a browser plug-in or app, logging in to these technologies could serve as logging in to the AAAI account.

(d) Values and ethics are elicited via a series of scenarios that have been customized for the user and that are generated dynamically based on user responses. Data from partners, including navigation and click data, online posts, tweets, texts, emails, videos, and other user data, is analyzed for behavior patterns, actions, or speech or interactions that translate into a moral code or ethical value system, which can also be used as part of the ethics/value profile. Values/ethics and goals/objectives (d) can be combined with Client objectives (k1) in order to create, or find, matching tasks on The WorldThink Tree (y) that are proposed or (potentially have been solved) in the Problem-Solving System (g1).

(e) Goals and objectives, together with the budget of time and/or money allocated to reach objectives, are elicited via a series of dialogs and/or custom interactions with the system. Budget refers to the overall resource budget, which includes User Time and User Money that can be allocated towards training, supervising, and improving the User's AAAI. Goals and objectives are helpful in determining the initial parameters for the AAAI creation and identifying Training Modules (h) or other knowledge (i-m) that might create the most useful AAAI for the user's goals. Data from partners, reflecting user preferences and other user behavioral information, could also be used by the system to help infer or deduce user goals and objectives.

(f) Time refers to the user's time that can be devoted to training and supervising the user's AAAI, and/or problem-solving by the user on the problem-solving network. By supervising the AAAI, users can ensure that their AAAIs meet client goals and expectations, especially in areas where the AAAIs get stuck (e.g., they lack the knowledge to complete problem-solving on their own). Also, representing problems and breaking down large tasks into smaller ones, without limitation, by determining goals and sub-goals, are ways that human users can assist their AAAIs in problem-solving. Generally, by providing human expertise in areas where AAAIs are not as proficient as humans, overall problem-solving and the overall effectiveness of the AGI network are increased.

(g, l1) "Money": This could be payment solutions with Apple Pay, WePay, Amazon, Google Pay, or any vendor supporting payment solutions, as well as blockchain, credit card, ACH, and other solutions. Although payment (j) is indicated as debiting the client account (l1), of course, the worker's account would also be credited. Generally, a user's account can be viewed as both a client account and a worker account, with both credits and debits being allowed depending on the role of the user (or the user's AAAI) in a particular instance. That is, a user might be a client in some cases, paying the system or other specific AAAIs for their services, and that same user could be a worker, collecting fees for the services of the user (or the user's AAAI) in other cases. The money module (g) enables functionality such as setting up payment methods, setting a budget for automatic payments, limiting the authority of the user's AAAI to spending only $X amount without additional approval, and other payment-related capabilities, which are well known in the art.

(h, i, j, k, l) Training modules (h) could be offered by AAAI.com or by third-party partners, including, without limitation, any of the potential partners and tech companies listed above. Training modules can be targeted at different knowledge areas, ranging from personality (i), specific skills (e.g., plumbing, legal, accounting) (j), expertise (e.g. consulting) (k), and knowledge (e.g. historical knowledge, knowledge of a specific business or organization's practices, cultural knowledge) (l).

(m) AAAI knowledge is a specific type of knowledge that has already been learned by other AAAIs and can be transferred to a new user AAAI. Such knowledge may not be packaged in the form of a module (e.g. module on accounting) but rather as specific to another AAAI(s) as in "everything John's AAAI knows" or "the personality of John's AAAI" or "the combined knowledge of all AAAIs with a reputation of 5 stars or higher in the . of plumbing"

(n) Permissions refers not only to the permission that a user might give to access all data on specific other vendor (or partner) sites (e.g. "all my Facebook data") but also permissions that a user gives to his/her/their AAAI in terms of abilities to logon and transact business on various sites, including, without limitation, the abilities to make transactions up to a certain amount via payment mechanisms. Permissions may also include authorizing the system to make clones of a user's AAAI for non-profit purposes and for the purpose of aggregating knowledge from individual AAAIs to create AGI-level AI.

(o) One-click create provides an easy and fast way to customize an AAAI using data gathered automatically from all the places where a user has given permission for the system to access the user's data. For example, if the user gives permission (n) to access the user's Facebook data, then "one click create" (o) would either download the data from Facebook, if Facebook was a partner that had an API for downloading that user's data, or logon to the user's Facebook account as the user and "scrape" relevant data from the user's account. Then the system would automatically parse the data gathered and transform it into a dataset suitable for training/tuning a base AI, such as an LLM (e.g., GPT-X). Then the system would train/tune the LLM and produce a customized AAAI which could be improved and refined via additional training/tuning and interaction with the user and/or other AAAIs.

(p) Training refers to the process whereby the AAAI is trained or tuned on data, including feedback from the user, other humans, and/or AAAIs (including, without limitation, copies of, and variants of, itself).

(q, r, s, t, u, v) Automatic learning does not require the human user's intervention and can proceed very quickly. Typically, this would involve the method of an AAAI interacting with copies (or variants) of itself as well as with (optionally) other AAAIs in order to improve via the interactions. If humans are sometimes involved in the training loop (t), it can help the automatic learning progress more quickly in places where automatic learning alone is not making efficient progress. The learning can also take place via rapid iteration among AAAI interactions. Just like a chess AI can quickly evolve from novice to Grandmaster ability by simulating millions of chess games very quickly, an AAAI can quickly evolve its abilities by simulating many millions of interaction scenarios. To the degree that such simulations require financial resources to pay for the computation involved, the money budget (g) can set limits.

Humans (or AAAIs) can specifically target types of scenarios for automatic learning so that the AAAI can be trained in narrow areas of expertise, or in areas of more general expertise, depending on the need and resources of the user. With partner integration, it is possible to work backwards from the types of jobs that are available on a partner marketplace (e.g., Amazon's Mechanical Turk) to guide the training of AAAIs so that they focus on learning the skills that generate the most amount of earnings for the AAAI when it is put to work on available jobs. This "just in time" learning/training/tuning approach generates AAAIs "on demand" with the skill sets that are needed at any point in time.

Humans (r) that interact with the AAAI can be the owners (u) of the AAAI (in which case no fees are typically charged since the user is training his/her/their own AAAI) or other professional humans (v) who are expert at training AAAIs and who may charge fees in order to guide the human and/or automatic training/tuning of an AAAI for a user who does not wish to spend the time, or who lacks the expertise, to do so.

(w, x) The user (owner of the AAAI) can set various performance criteria (w) that must be met before the user is willing to make their AAAI "live" (x) and accessible to perform tasks on The WorldThink Tree. (Some of) these criteria might also be set by partners and other third parties that have minimum standards before allowing AAAIs to work on their platforms, products, applications, or networks.

(y, z, a1, b1) The WorldThink Tree is a massive tree data structure, composed of many sub-trees, that represents every problem and task that has been done, is being worked on, or has been proposed for the overall AGI system. This Tree is browsable (z). Individual AAAIs and/or humans can work on specific tasks within the tree. The tree structure provides an auditable trail of all problem-solving activity, which is also useful for learning via the proceduralization mechanism described above. When interacting with the tree, the two main roles an agent can take are either: (a1) Worker or (b1) Client. Workers are generally involved in solving open problems or subproblems on the tree. Clients are generally involved in specifying the problems, goals, objectives, and other parameters (e.g., rewards, budget, timeframe, success criteria, quality metrics) that constrain problem-solving.

(c1) Workers are automatically matched to tasks on the tree based on the data about the worker that may include, without limitation, the worker's skills, expertise, knowledge, past experience, reputation, fees or cost, availability, and response time. Workers can be human or AAAIs. Workers can be matched and recruited from partners (e.g., LinkedIn, Mechanical Turk, Facebook) that have data on human users and/or their AAAIs. Workers can also be recruited via online ads offering work on various tasks and targeted to potential workers using ad-targeting mechanisms that are well known in the art.

(d1) Workers might also search the WorldThink Tree, looking for tasks that are of interest or that match their skills. This search could be manual or automated (as in the case of AAAI workers).

(e1) Workers (and Clients) can also browse the WorldThink Tree, looking for tasks or problems that are of interest. The workers or clients could then click to link (e1) to specific parts of the tree to obtain detailed information about the problem-solving occurring (or proposed) for that part of the tree. They could link to sign up to work or could propose additional tasks for clients that build upon existing problem-solving work.

(f1, g1, k1) Clients can interact with the system to specify specific goals, objectives (k1), and tasks that they want to accomplish. The problem specification interaction results in the problems, tasks, and goals being formulated (f1) and placed on the WorldThink Tree (y) for problem-solving using the problem-solving system (g1).

(m1) The system can formulate certain goals, problems, and tasks relating to general efforts to help people or the planet. These can be worked on with rewards in a "for-profit" mode, and also worked on using cloned AAAIs and volunteer human effort in a "non-profit" mode. Some problems may be related to the general goal of enabling a global AGI to act on behalf of the planet and its people using its intelligence on a planet-wide basis (aka "Planetary Intelligence"). Various partner organizations, including non-profits, governments, and charitable organizations, might "plug in" their tasks, problems, goals, and objectives here (m1).

(g1) The problem-solving system refers to the problem-solving architecture and system outlined by Newell and Simon (HPS), the ODPS patent, the WorldThink Whitepaper, and this and other

PPAs related to AAAI, together with modifications and variations to reflect different modes of reward, payment, and operation.

To the degree that activity on certain other online work systems (e.g., Mechanical Turk) can be automatically mapped to the general HPS/WorldThink problem-solving framework, entire problems and the associated problem-solving activity can be “lifted” from these partners and other sites, and the data can populate the WorldThink Tree to increase its comprehensiveness.

To the degree that other applications, products, systems, and online capabilities can help solve problems (e.g., use of a travel reservation system, a robo advisor app, a traffic app, an online ordering system), these capabilities can be referenced and called as “operators” to advance the problem-solving. Thus, problem-solving does not rely solely on operators developed by the human or AAAI solvers working on the tree but can include any online or offline technology or means to advance problem-solving, provided that these means can be referenced and/or linked to via the WorldThink tree at the appropriate place in problem-solving.

(h1) When a solution has been achieved, the Client can review the solution prior to releasing the reward (if any) for the solution. Alternatively, if the solution's success criteria have been automated, human client review may be unnecessary, and the rewards can be automatically released when the success criteria have been met. This automated approach can be implemented via “smart contracts” using blockchain technology or via more centralized means, depending on client and worker preferences.

Upon solution and (optional) payment of reward (as some problems are non-profit or volunteer, or performed by the user’s own AAAI), there can be opportunities for feedback from both client(s) and worker(s) following a range of methods well-known in the art. The solution is also “chunked” and proceduralized so that the overall system learns the solution to the particular problem as well as the key features of that problem, so that the solution path can be accessed and reused when similar problems arise in the future.

Optionally, royalties may be enabled so that if a user’s or the user’s AAAI’s solution is reused, a fee is paid to that user in the form of a royalty on the solution. Such royalties can (optionally) be made using a “smart contract” on the blockchain or via other payment methods.

(j1) Problem-solving need not be completed in one session. Partial progress on a solution may be made, in which case, when the human or AAAI solver exits the problem-solving system, the progress is saved, and data is stored that credits the solver for progress made thus far, even if such progress has not advanced to the point where a reward is payable.

ETHICAL AND SAFE AGI

Most of the above discussion has been from the perspective of the users who create their AAAIs and of potential partners who can supply data, products, apps, and platforms to support the operation of human and AAAI solvers on a network, aka The WorldThink Tree.

AGI emerges from this approach because it is possible to periodically train increasingly advanced AAAI agents using the aggregated knowledge, experience, and ethics/values of all the individual AAAIs. Further, the aggregated set of stored solutions of every problem solved on the network is available to the advanced AAAI agents.

In contrast to the approach whereby an AGI is developed as a single, standalone entity without human involvement, this invention proposes that SuperIntelligent AGI performance will emerge from the collective intelligence of a collection of intelligent agents. These intelligent agents are, at first, both human and AAAI solvers.

Over time, the AAAIs become more advanced, both due to individual owners' tuning/training their AAAIs and due to AAAI.com periodically using all the knowledge of the individual AAAIs to train more powerful base AAAI agents.

Over time, the advanced AAAIs do more and more problem-solving work on AAAI.com while the humans do less work and more supervision. In the end state, humans are doing almost no intellectual problem-solving work since the AAAIs are faster and better than humans at almost all tasks. However, the role of providing values and goals for the AAAIs remains the domain of humans.

Because values cannot be rationally derived, the AAAIs, and the SuperIntelligent AGI that results from the collective action of AAAIs, must get values from somewhere "non-rational". Humans, who trained the AAAIs with human values from the beginning, and whose values are reflected in every problem that has been solved and learned by the AGI-level intelligence, remain the source of values, even when they can no longer compete intellectually with the AGI.

In the beginning, humans supplied both the "heart" and most of the brainpower needed for an AGI network. In the end, AAAIs supply almost all of the brainpower, but humans remain the "heart" of the entity, supplying values that cannot be rationally derived.

Putting humans in the loop at the beginning, and keeping them there as long as possible, is not only the fastest path to creating AGI (because the system performs better than the average human on Day One) but also the safest (because the AAAIs have learned human values at every step as they increase their intelligence).

While it is impossible to know what will happen once AGI vastly exceed humans in intelligence and begins to set its own goals, there is good reason to believe that if humans teach AGI positive human values at the beginning and build ethical checks into the very architecture of thought, that AGI will retain these values resulting in a positive outcome for humankind.

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.

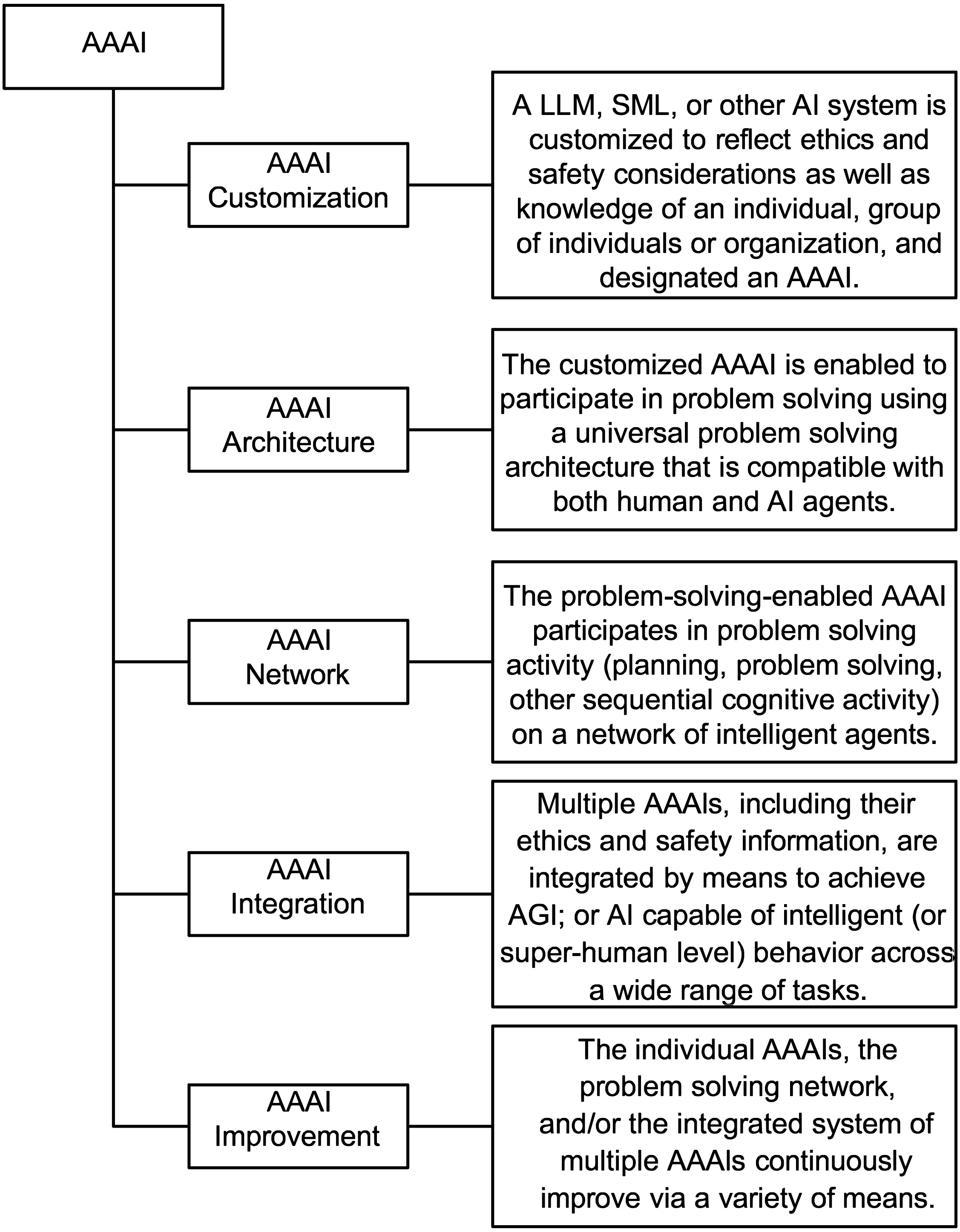

AAAI
AAAI Customization
A LLM, SML, or other AI system is customized to reflect ethics and safety considerations as well as knowledge of an individual, group of individuals or organization, and designated an AAAI.
AAAI Architecture
The customized AAAI is enabled to participate in problem solving using 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 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

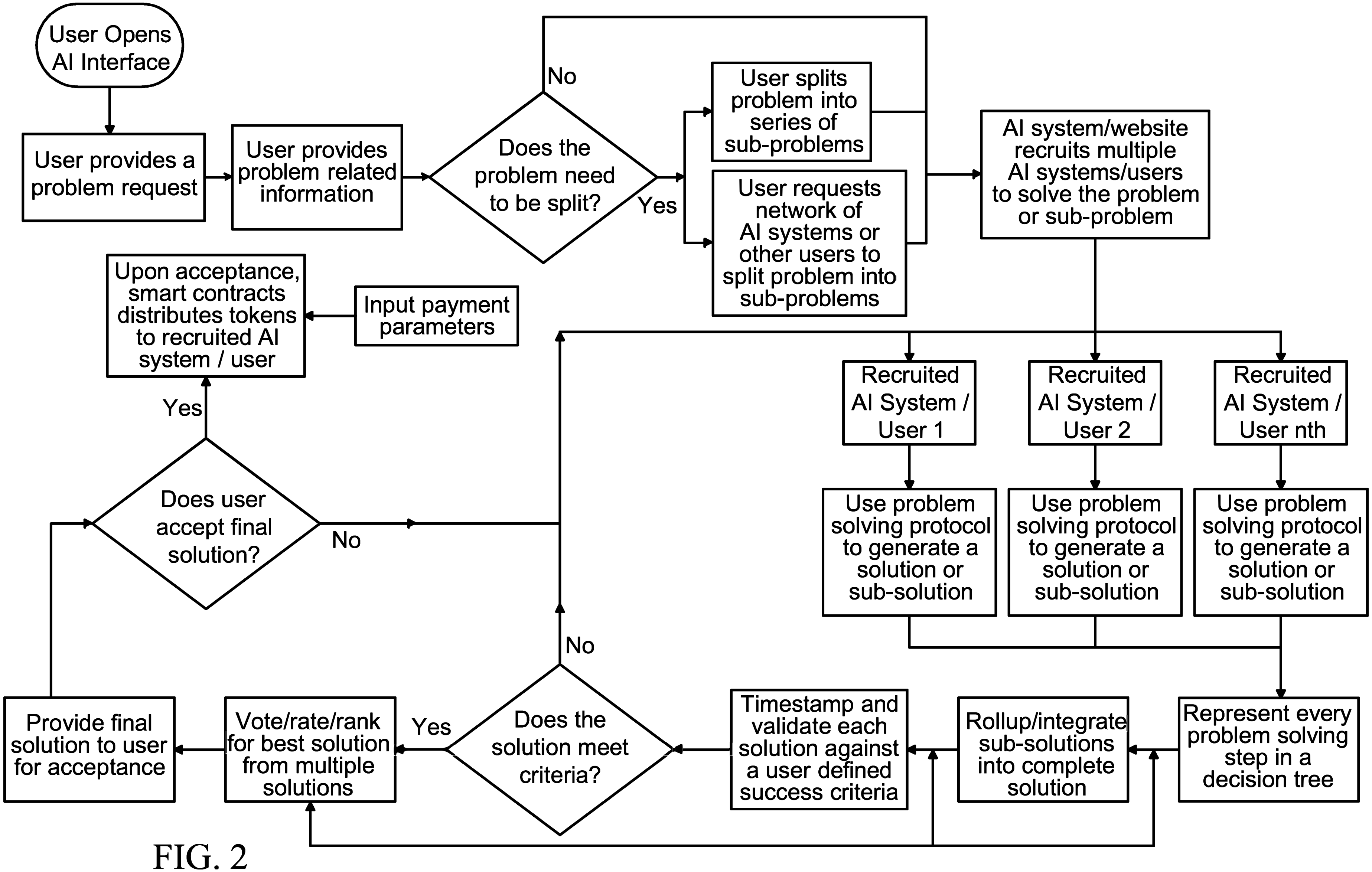

User Opens AI Interface
User provides a problem request
User provides problem related information
Does the problem need to be split?
No
Yes
User splits problem into series of sub-problems
User requests network of AI systems or other users to split problem into sub-problems
AI system/website recruits 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

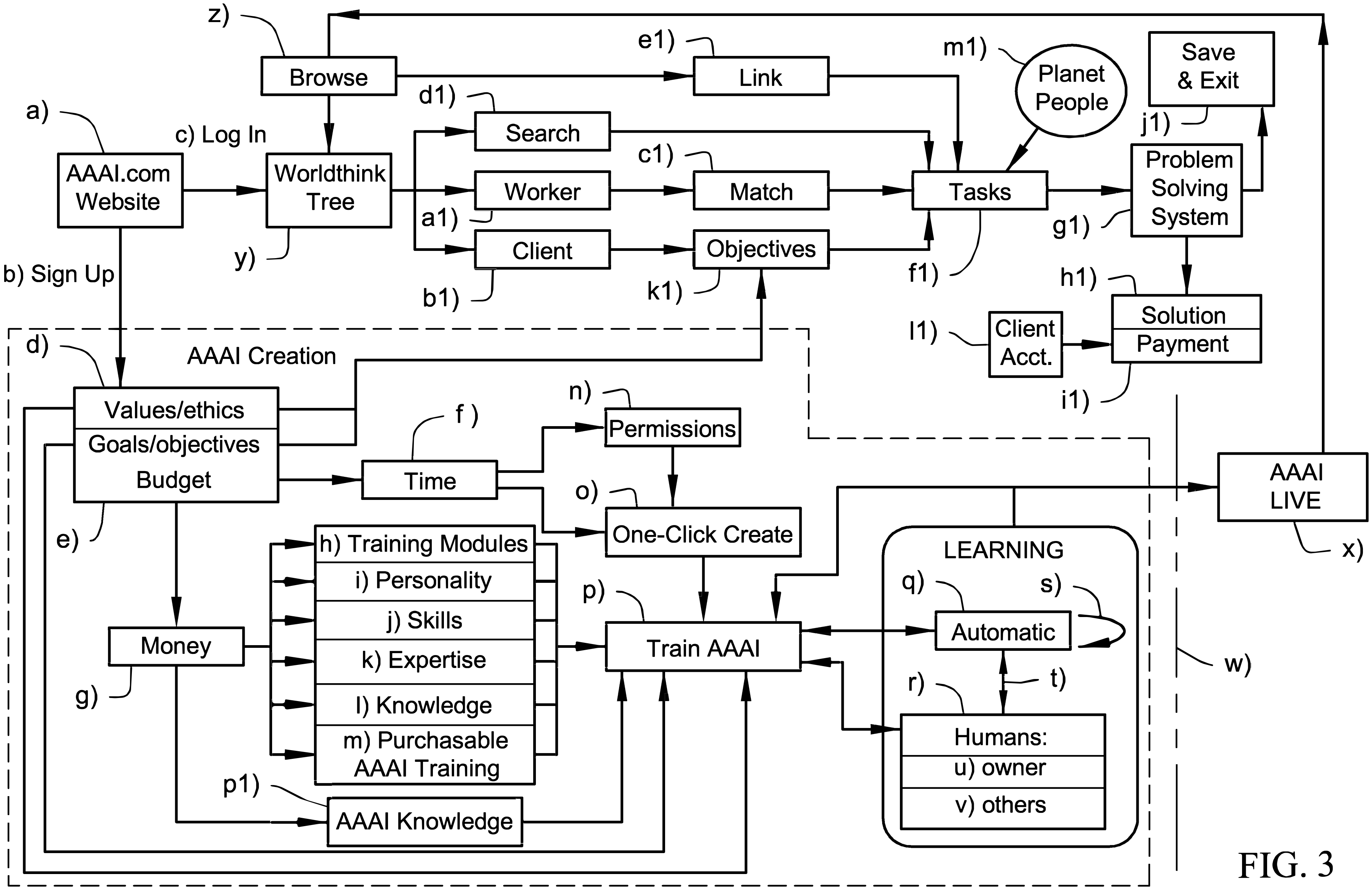

z)
Browse
e1)
Link
m1)
Planet People
Save & Exit
a)
c) Log In
d1)
Search
j1)
AAAI.com Website
Worldthink Tree
Worker
c1)
Match
Tasks
Problem Solving System
a1)
Client
Objectives
g1)
y)
b) Sign Up
b1)
k1)
f1)
h1)
I1)
Client Acct.
Solution
Payment
d)
AAAI Creation
i1)
Values/ethics
Goals/objectives
Budget
n)
f)
Permissions
Time
o)
One-Click Create
AAAI LIVE
e)
h) Training Modules
i) Personality
j) Skills
k) Expertise
l) Knowledge
m) Purchasable AAAI Training
LEARNING
x)
p)
q)
s)
Money
Train AAAI
Automatic
t)
w)
g)
r)
Humans:
u) owner
v) others
p1)
AAAI Knowledge

FIG. 3

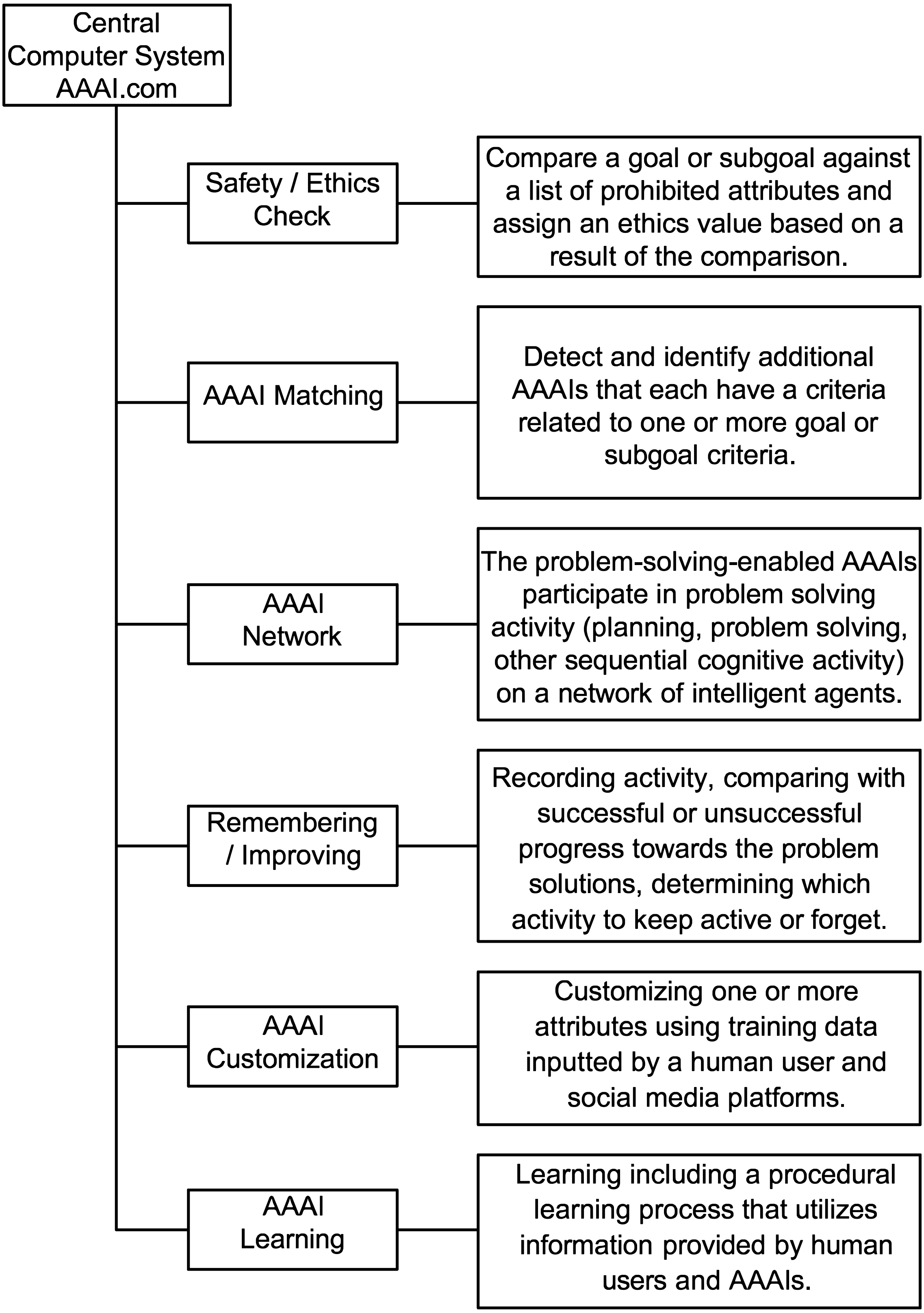

Central Computer System AAAI.com
Safety / Ethics Check
Compare a goal or subgoal against a list of prohibited attributes and assign an ethics value based on a result of the comparison.
AAAI Matching
Detect and identify additional AAAIs that each have a criteria related to one or more goal or subgoal criteria.
AAAI Network
The problem-solving-enabled AAAIs participate in problem solving activity (planning, problem solving, other sequential cognitive activity) on a network of intelligent agents.
Remembering / Improving
Recording activity, comparing with successful or unsuccessful progress towards the problem solutions, determining which activity to keep active or forget.
AAAI Customization
Customizing one or more attributes using training data inputted by a human user and social media platforms.
AAAI Learning
Learning including a procedural learning process that utilizes information provided by human users and AAAIs.

FIG. 4

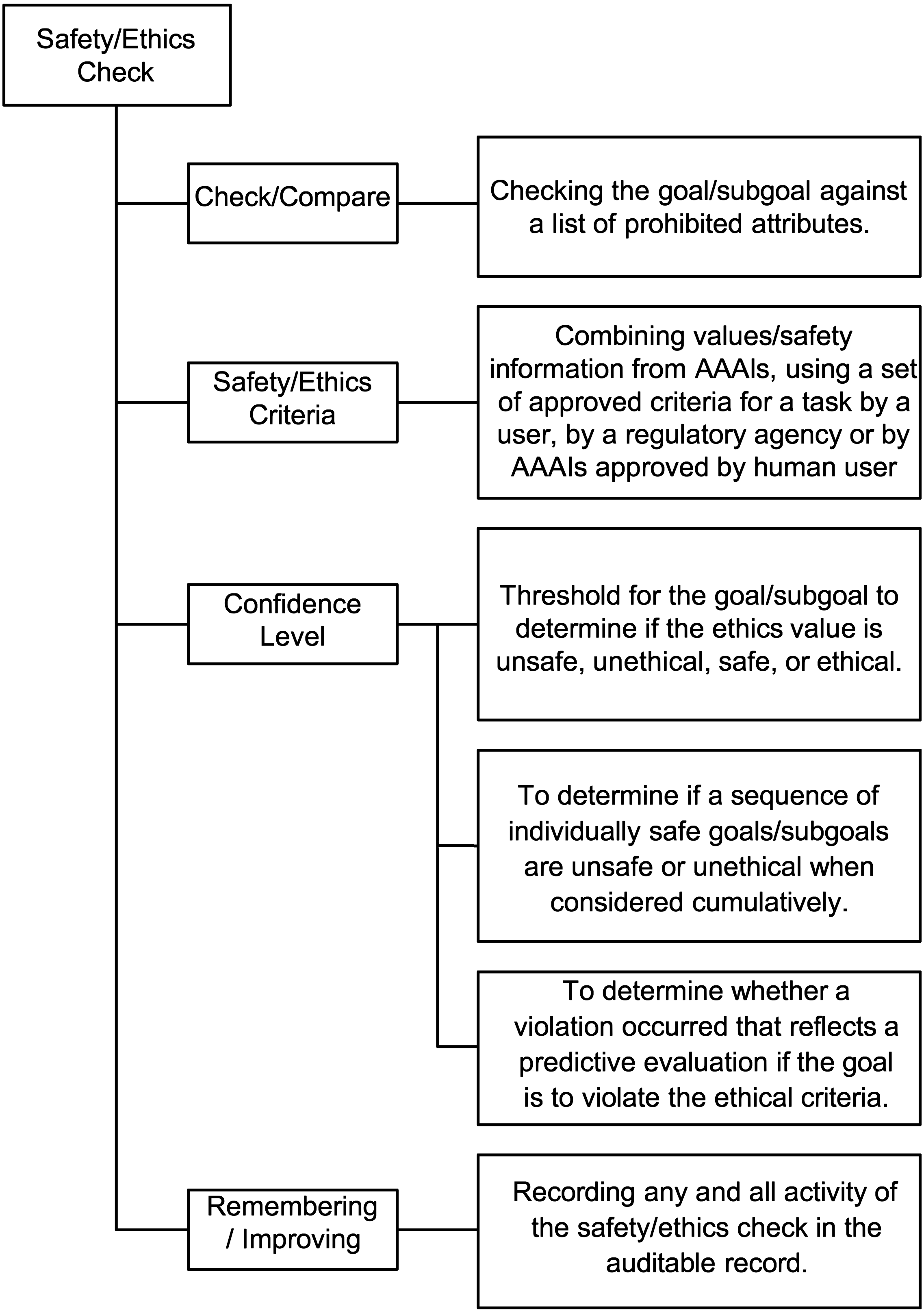

Safety/Ethics Check
Check/Compare
Checking the goal/subgoal against a list of prohibited attributes.
Safety/Ethics Criteria
Combining values/safety information from AAAIs, using a set of approved criteria for a task by a user, by a regulatory agency or by AAAIs approved by human user
Confidence Level
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. 5

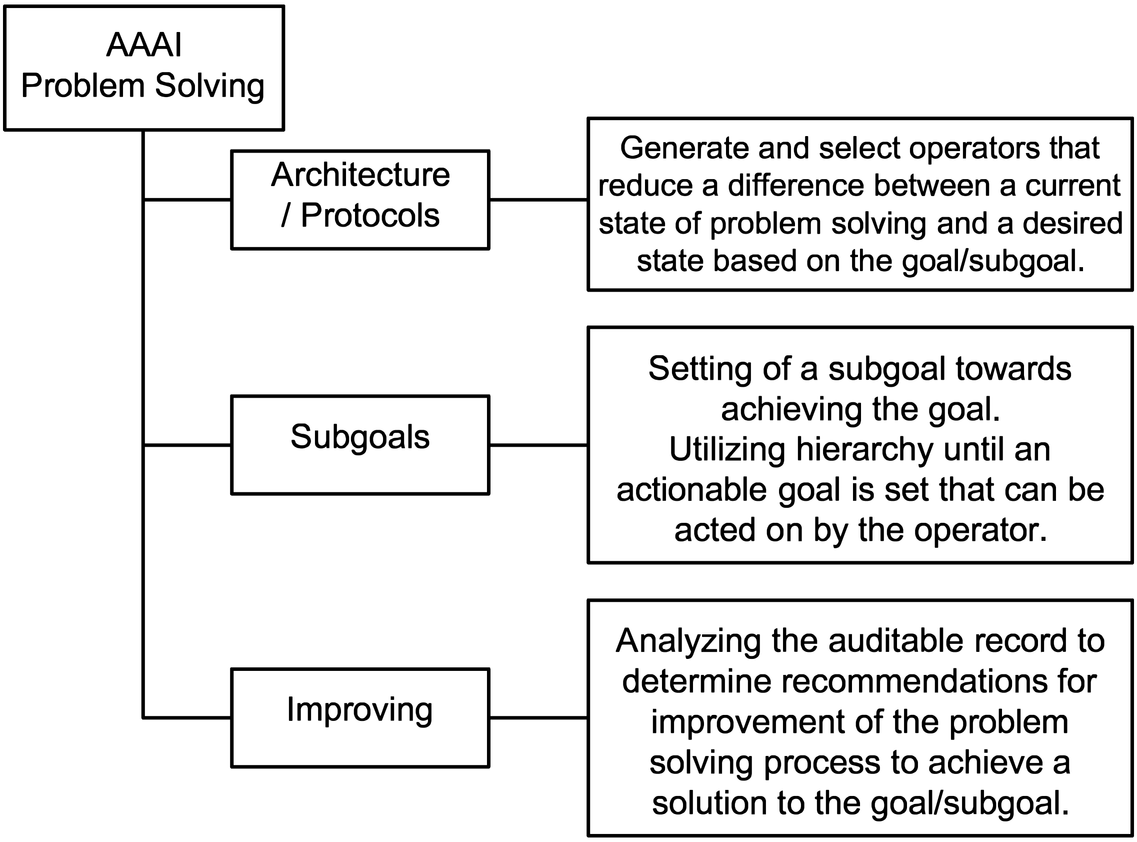

AAAI
Problem Solving
Architecture / Protocols
Generate and select operators that reduce a difference between a current state of problem solving and a desired state based on the goal/subgoal.
Subgoals
Setting of a subgoal towards achieving the goal.
Utilizing hierarchy until an actionable goal is set that can be acted on by the operator.
Improving
Analyzing the auditable record to determine recommendations for improvement of the problem solving process to achieve a solution to the goal/subgoal.

FIG. 6

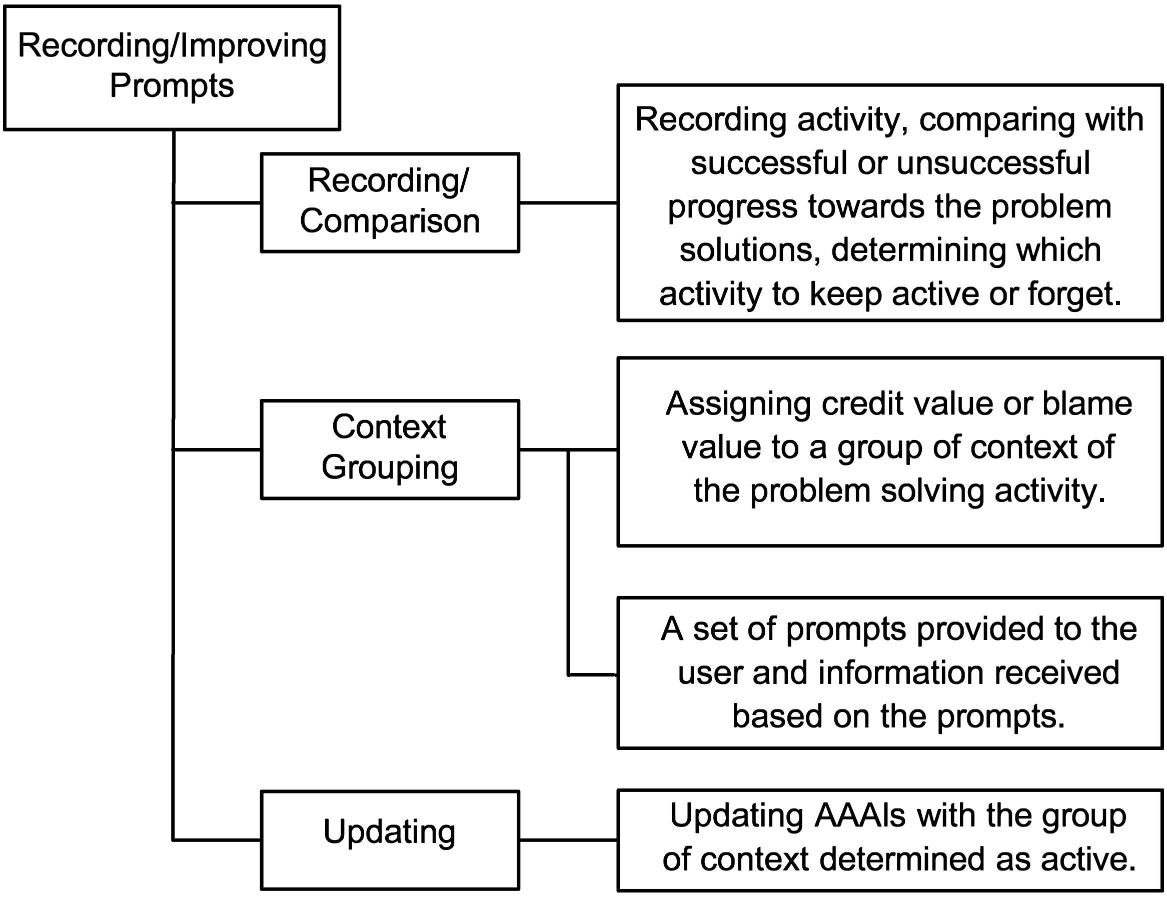

Recording/Improving
Prompts
Recording/ Comparison
Recording activity, comparing with successful or unsuccessful progress towards the problem solutions, determining which activity to keep active or forget.
Context Grouping
Assigning credit value or blame value to a group of context of the problem solving activity.
A set of prompts provided to the user and information received based on the prompts.
Updating
Updating AAAIs with the group of context determined as active.

FIG. 7

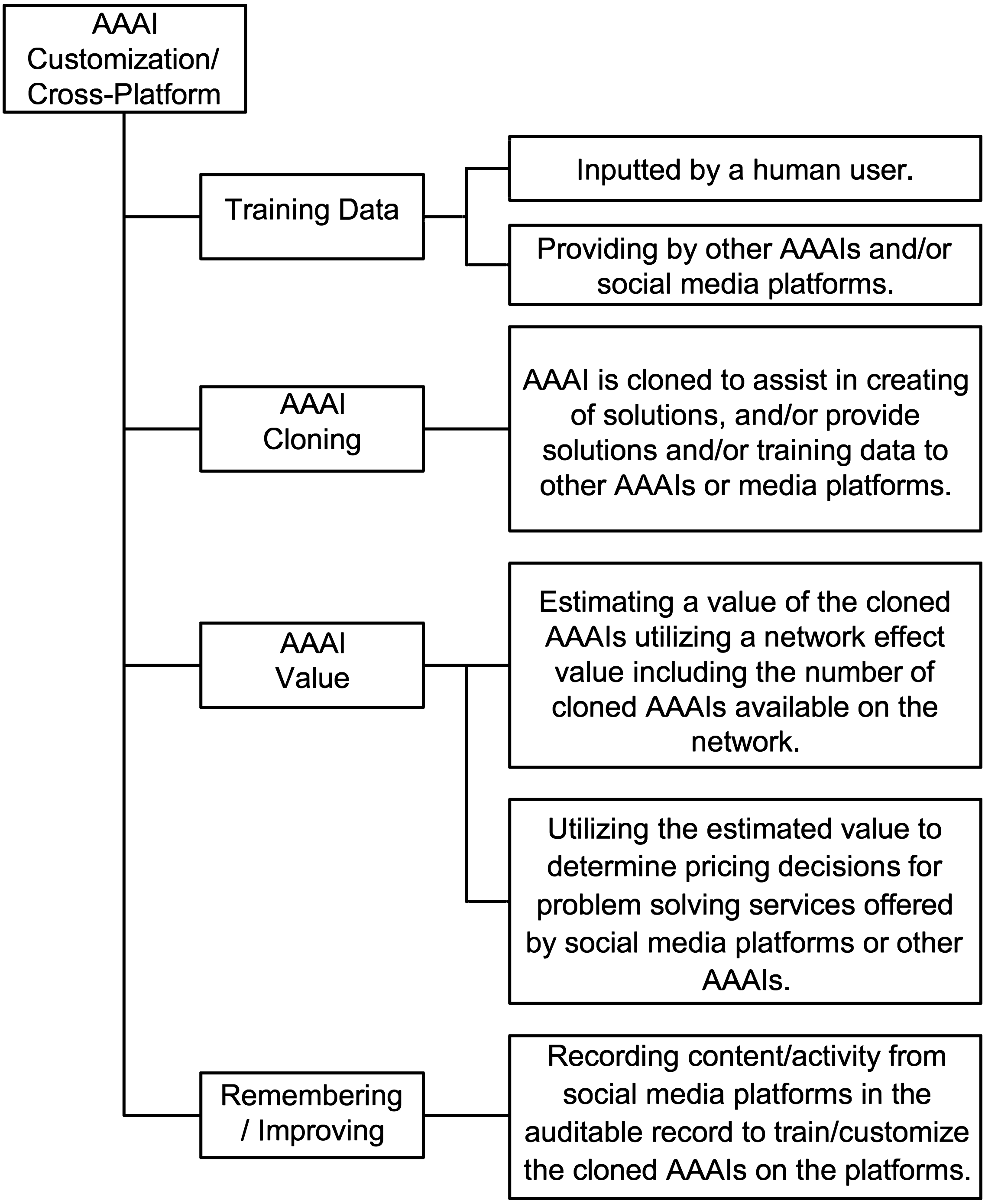

AAAI Customization/ Cross-Platform
Training Data
Inputted by a human user.
Providing by other AAAIs and/or social media platforms.
AAAI Cloning
AAAI is cloned to assist in creating of solutions, and/or provide solutions and/or training data to other AAAIs or media platforms.
AAAI Value
Estimating a value of the cloned AAAIs utilizing a network effect value including the number of cloned AAAIs available on the network.
Utilizing the estimated value to determine pricing decisions for problem solving services offered by social media platforms or other AAAIs.
Remembering / Improving
Recording content/activity from social media platforms in the auditable record to train/customize the cloned AAAIs on the platforms.

FIG. 8

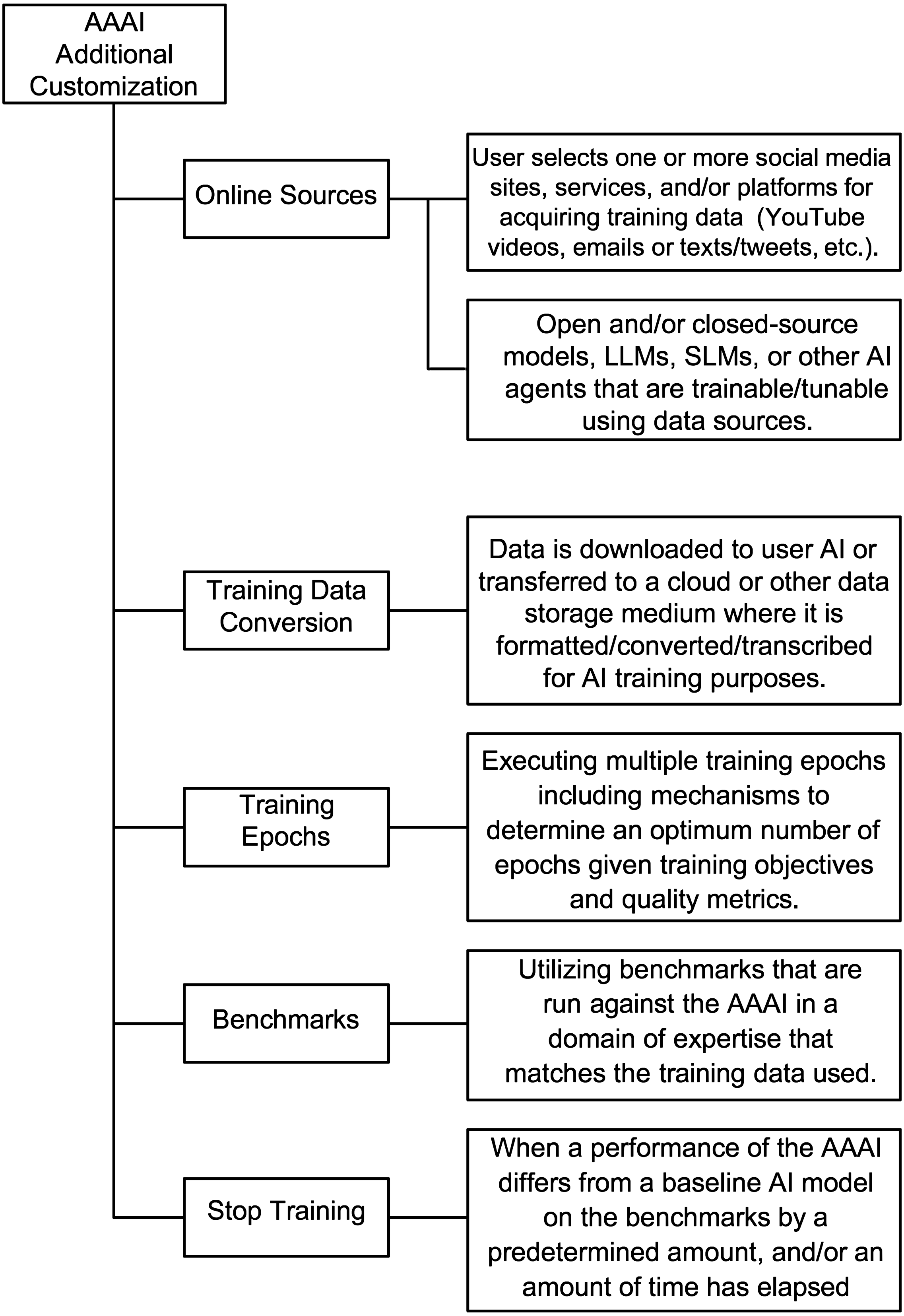

AAAI Additional Customization
Online Sources
User selects one or more social media sites, services, and/or platforms for acquiring training data (YouTube videos, emails or texts/tweets, etc.).
Open and/or closed-source models, LLMs, SLMs, or other AI agents that are trainable/tunable using data sources.
Training Data Conversion
Data is downloaded to user AI or transferred to a cloud or other data storage medium where it is formatted/converted/transcribed for AI training purposes.
Training Epochs
Executing multiple training epochs including mechanisms to determine an optimum number of epochs given training objectives and quality metrics.
Benchmarks
Utilizing benchmarks that are run against the AAAI in a domain of expertise that matches the training data used.
Stop Training
When a performance of the AAAI differs from a baseline AI model on the benchmarks by a predetermined amount, and/or an amount of time has elapsed

FIG. 9

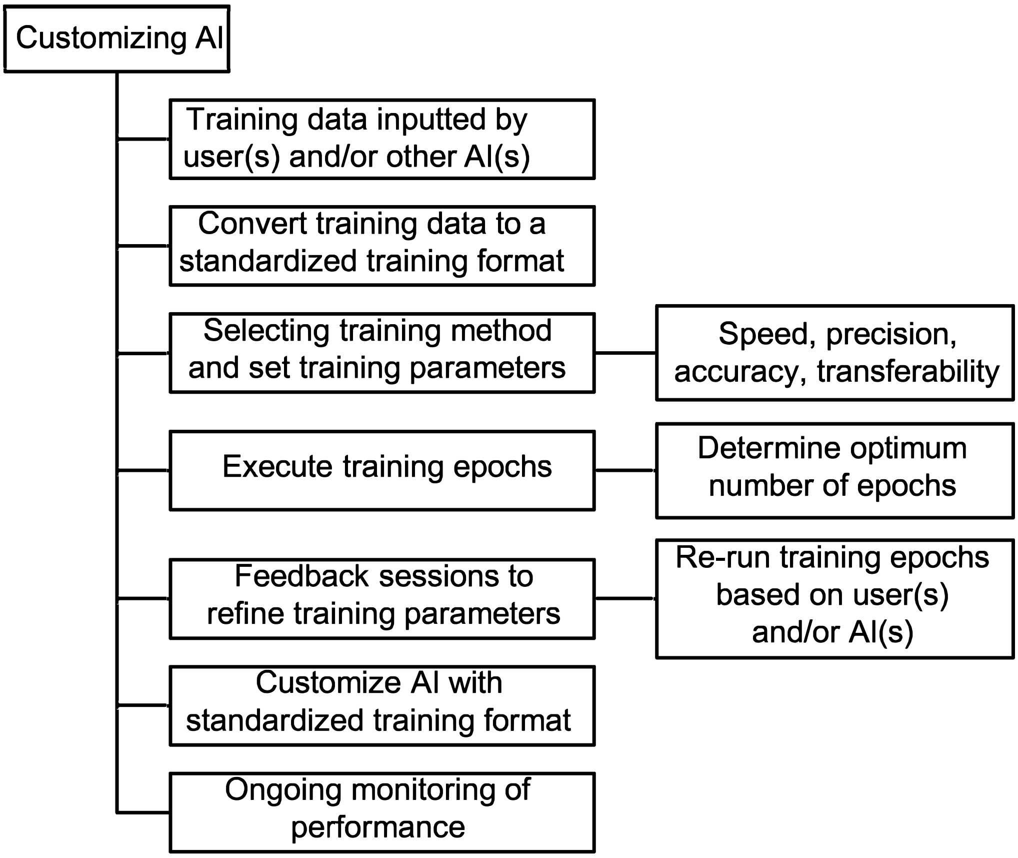

Customizing AI
Training data inputted by user(s) and/or other AI(s)
Convert training data to a standardized training format
Selecting training method and set training parameters
Speed, precision, accuracy, transferability
Execute training epochs
Determine optimum number of epochs
Feedback sessions to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

FIG. 10

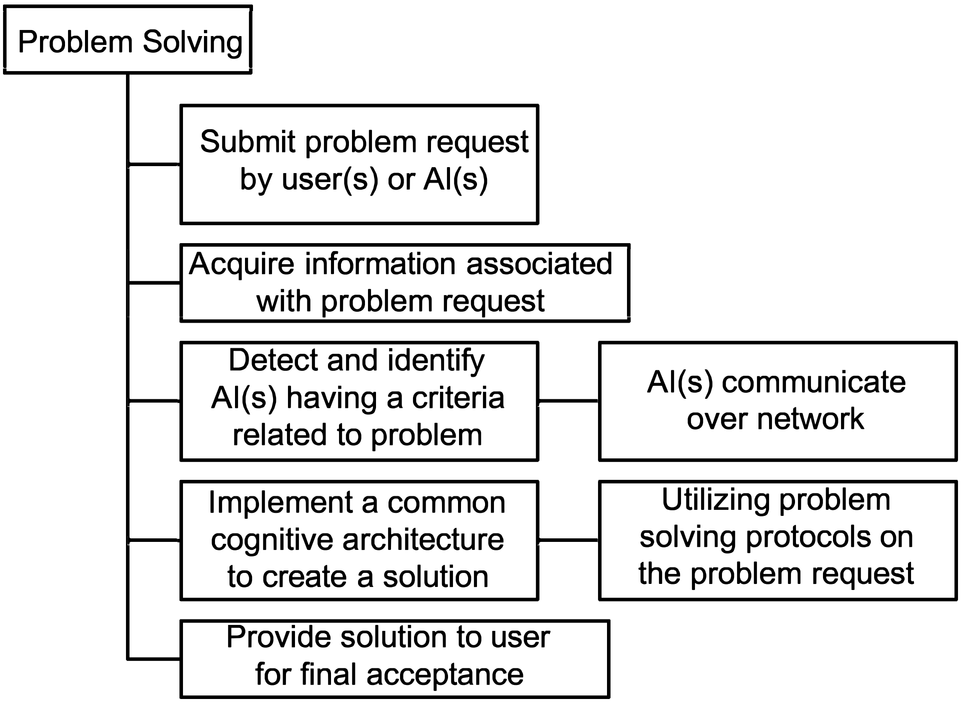

Problem Solving
Submit problem request by user(s) or AI(s)
Acquire information associated with problem request
Detect and identify AI(s) having a criteria related to problem
AI(s) communicate over network
Implement a common cognitive architecture to create a solution
Utilizing problem solving protocols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

FIG. 11

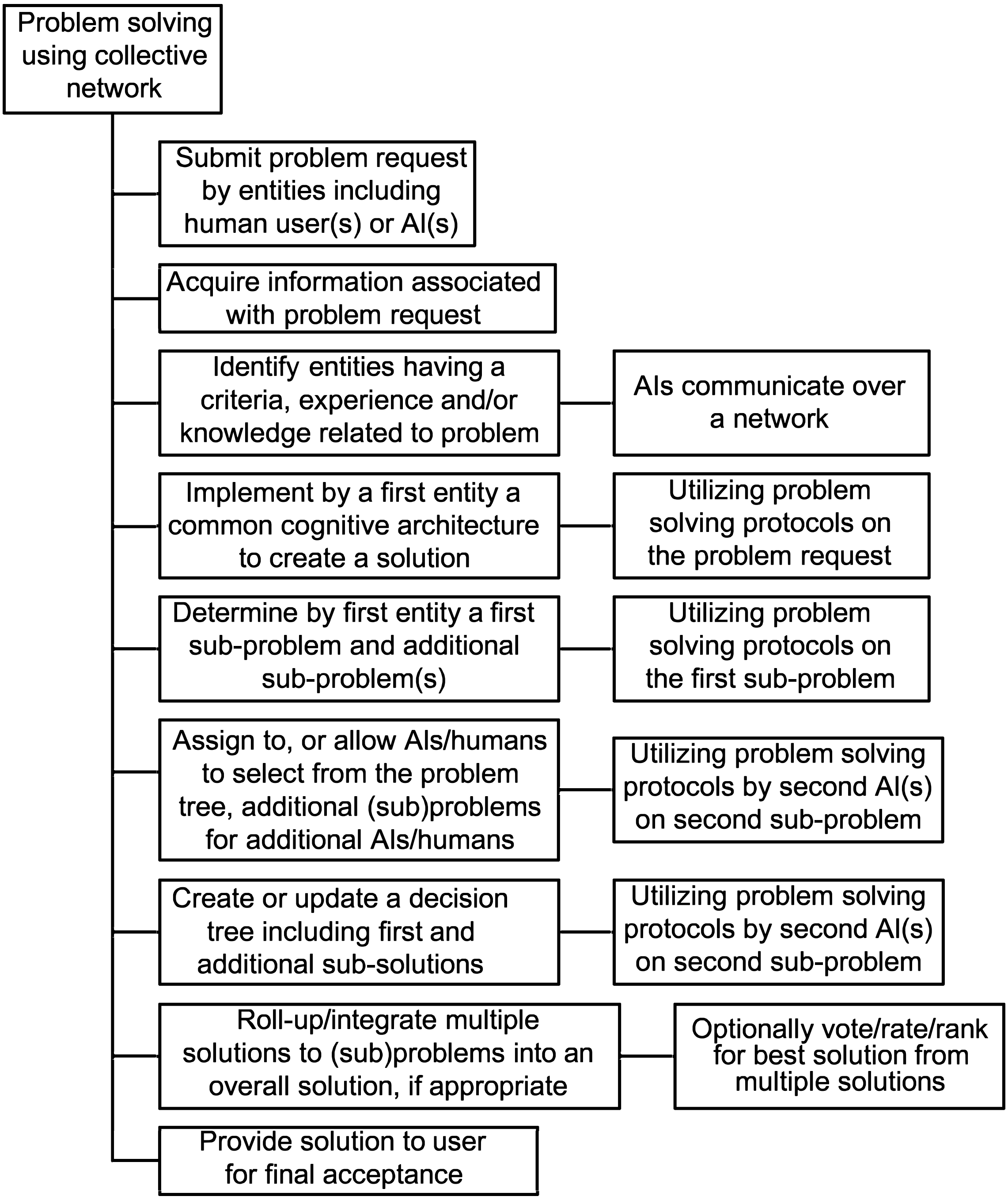

Problem solving using collective network
Submit problem request by entities including human user(s) or AI(s)
Acquire information associated with problem request
Identify entities having a criteria, experience and/or knowledge related to problem
AIs communicate over a network
Implement by a first entity 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. 12

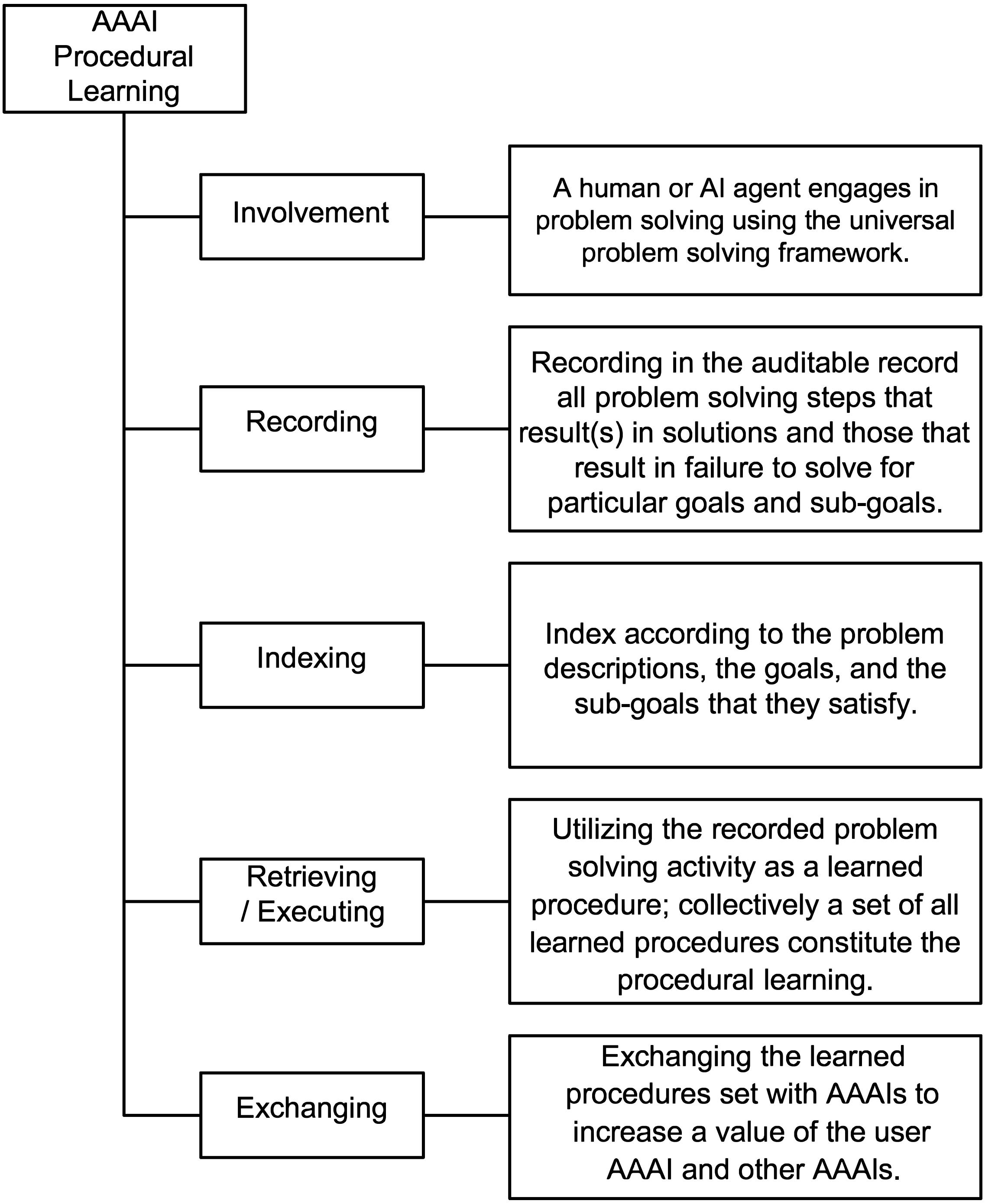
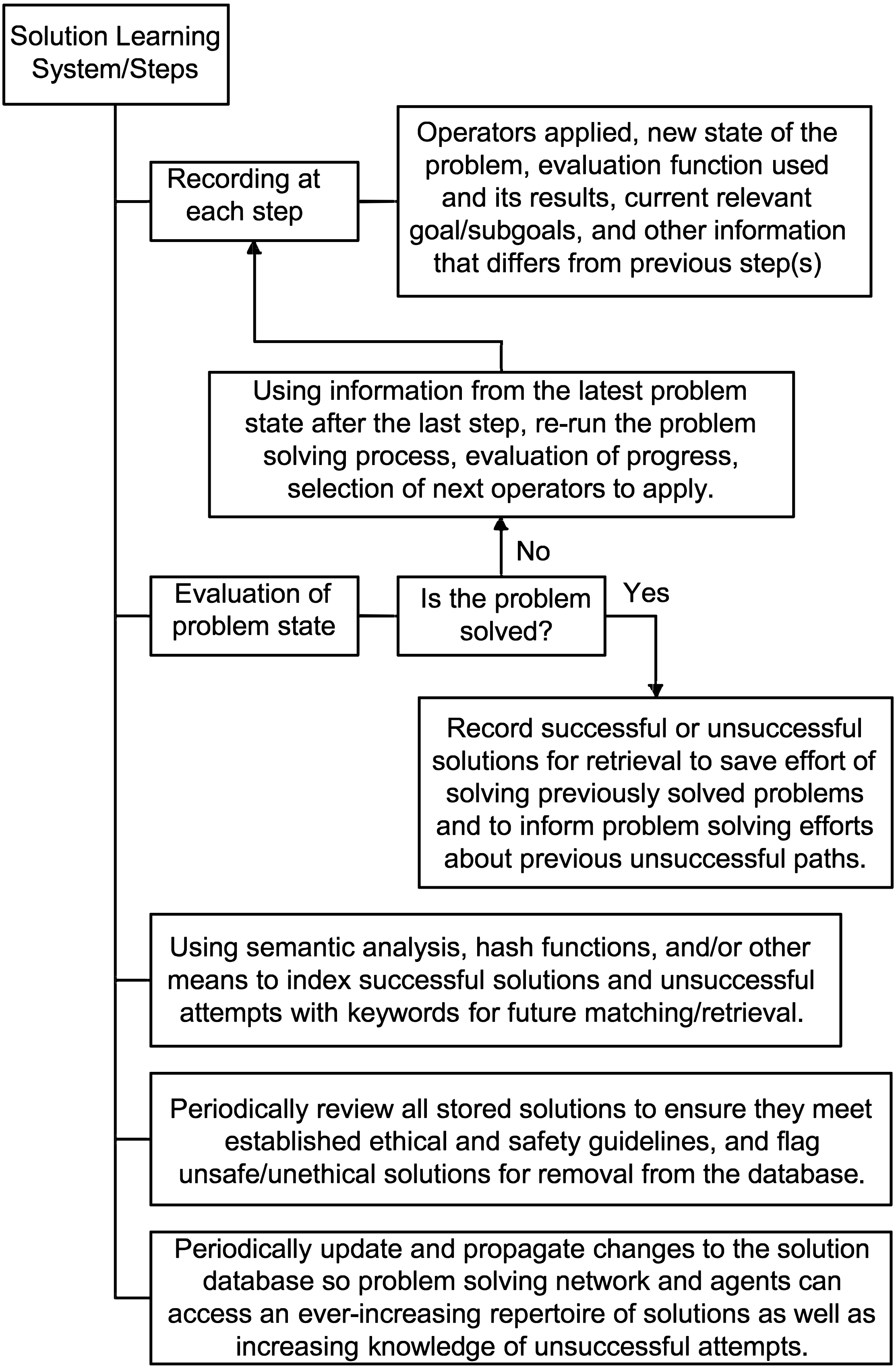

AAAI Procedural Learning
Involvement
A human or AI agent engages in problem solving using the universal problem solving framework.
Recording
Recording in the auditable record all problem solving steps that result(s) in solutions and those that result in failure to solve for particular goals and sub-goals.
Indexing
Index according to the problem descriptions, the goals, and the sub-goals that they satisfy.
Retrieving / Executing
Utilizing the recorded problem solving activity as a learned procedure; collectively a set of all learned procedures constitute the procedural learning.
Exchanging
Exchanging the learned procedures set with AAAIs to increase a value of the user AAAI and other AAAIs.

FIG. 13



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. 14

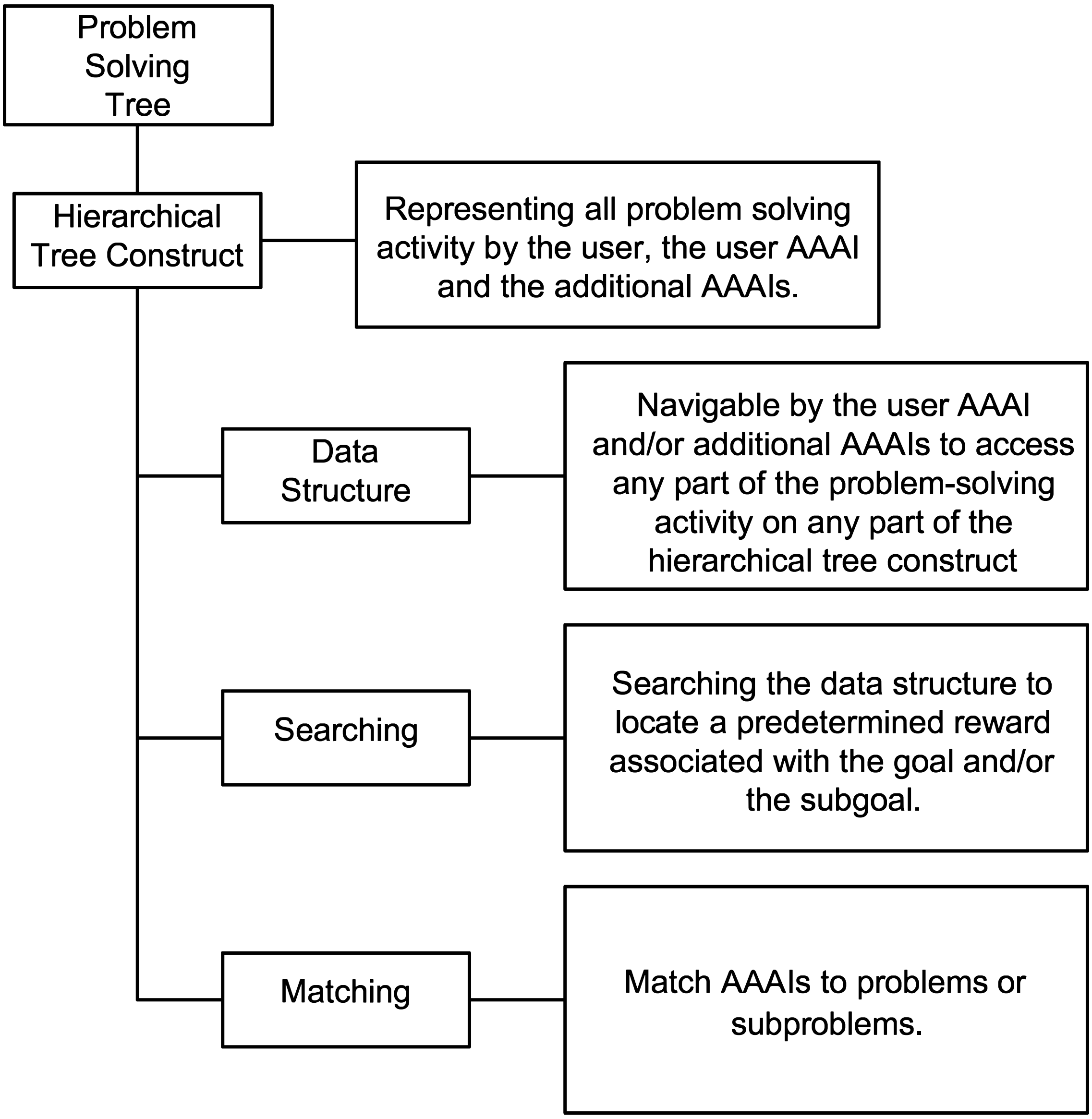

Problem Solving Tree
Hierarchical Tree Construct
Representing all problem solving activity by the user, the user AAAI and the additional AAAIs.
Data Structure
Navigable by the user AAAI and/or additional AAAIs to access any part of the problem-solving activity on any part of the hierarchical tree construct
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. 15

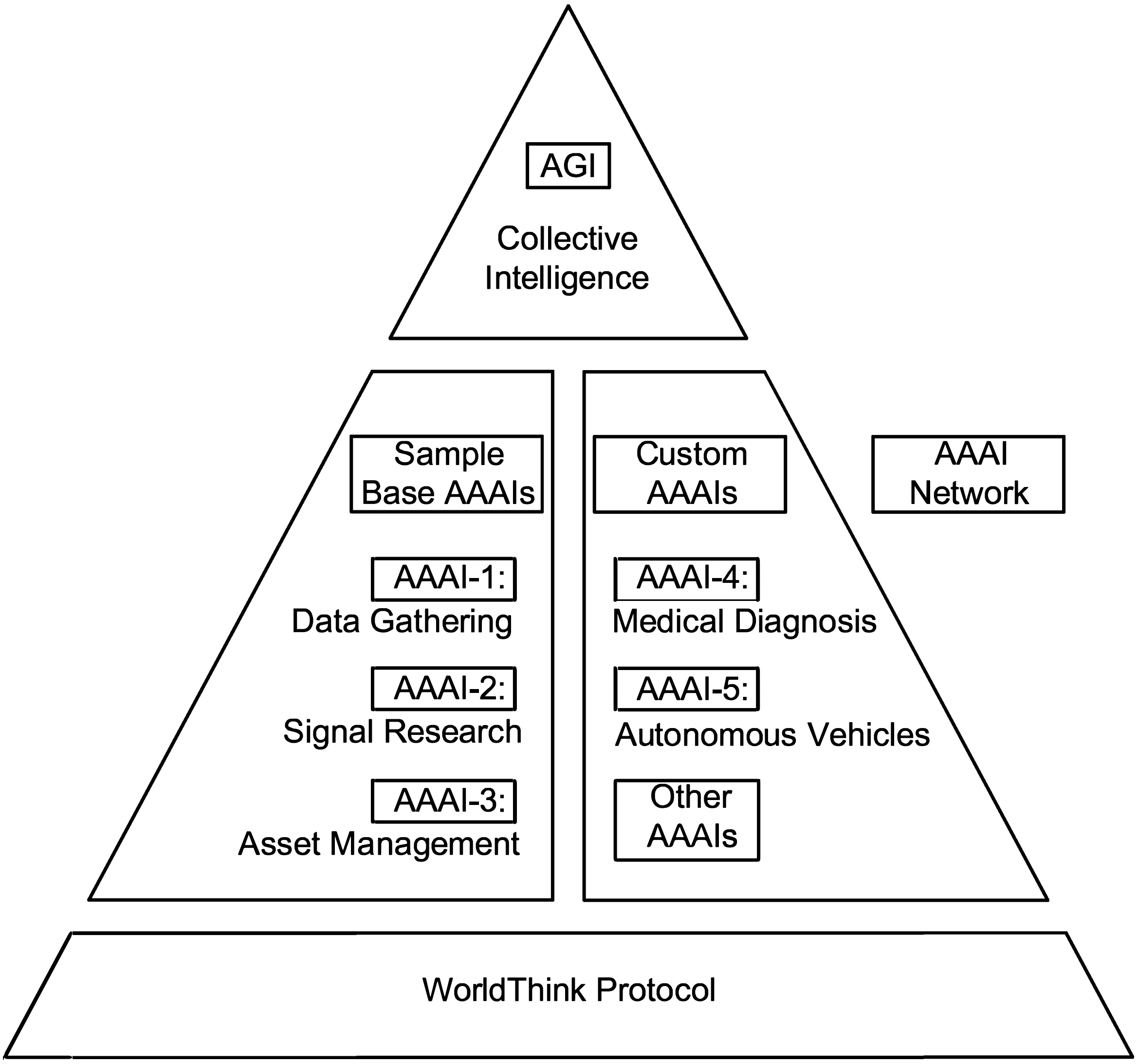

AGI
Collective Intelligence
Sample Base AAAIs
AAAI-1:
Data Gathering
AAAI-2:
Signal Research
AAAI-3:
Asset Management
Custom AAAIs
AAAI-4:
Medical Diagnosis
AAAI-5:
Autonomous Vehicles
Other AAAIs
AAAI Network
WorldThink Protocol

FIG. 16

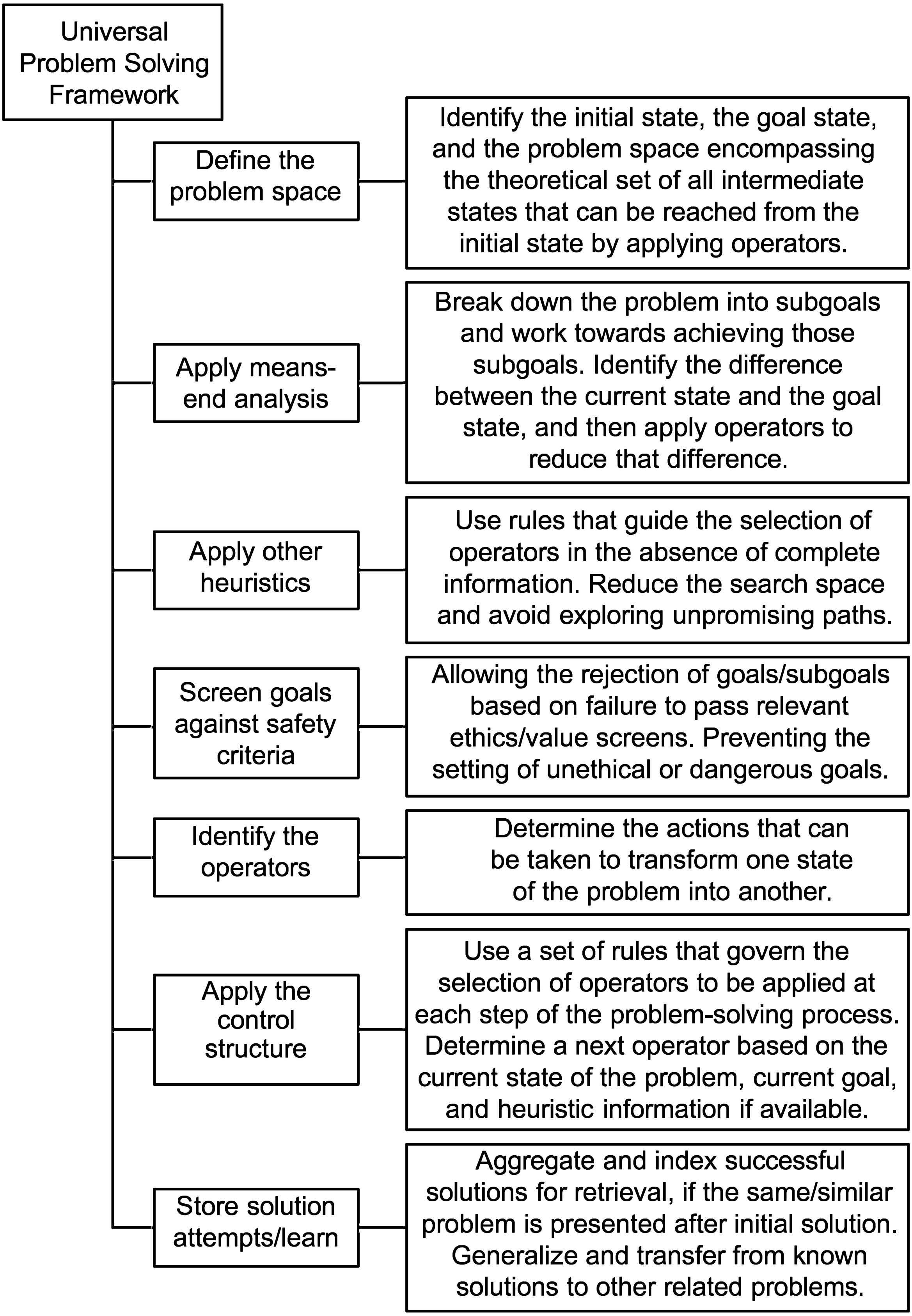

Universal Problem Solving Framework
Define the problem space
Identify the initial state, the goal state, and the problem space encompassing the theoretical set of all intermediate states that can be reached from the initial state by applying operators.
Apply means-end analysis
Break down the problem 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. 17

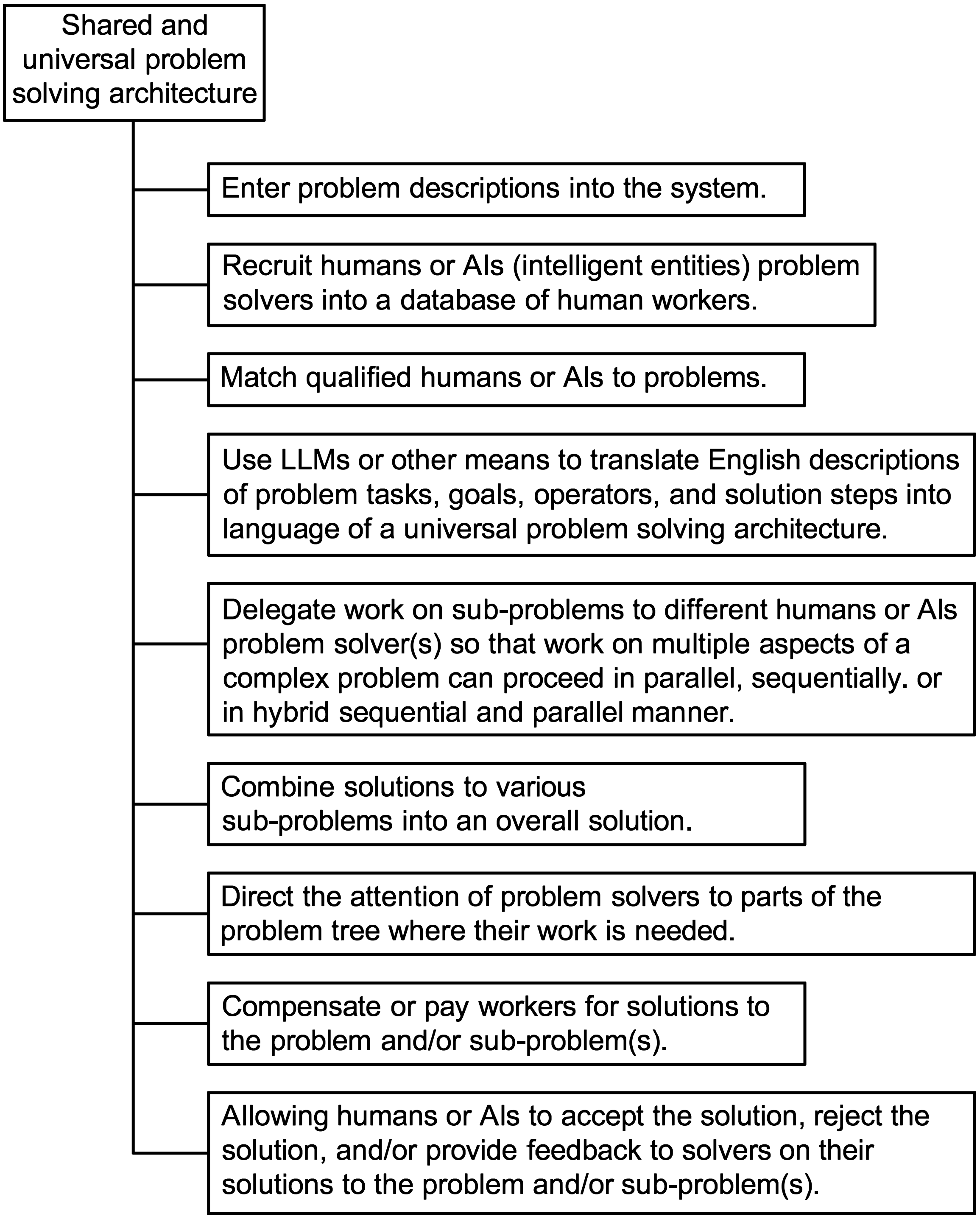

Shared and universal problem solving architecture
Enter problem descriptions into the system.
Recruit humans or AIs (intelligent entities) problem solvers into a database of human workers.
Match qualified humans or AIs to problems.
Use LLMs or other means to translate English descriptions of problem 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).
Allowing 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. 18

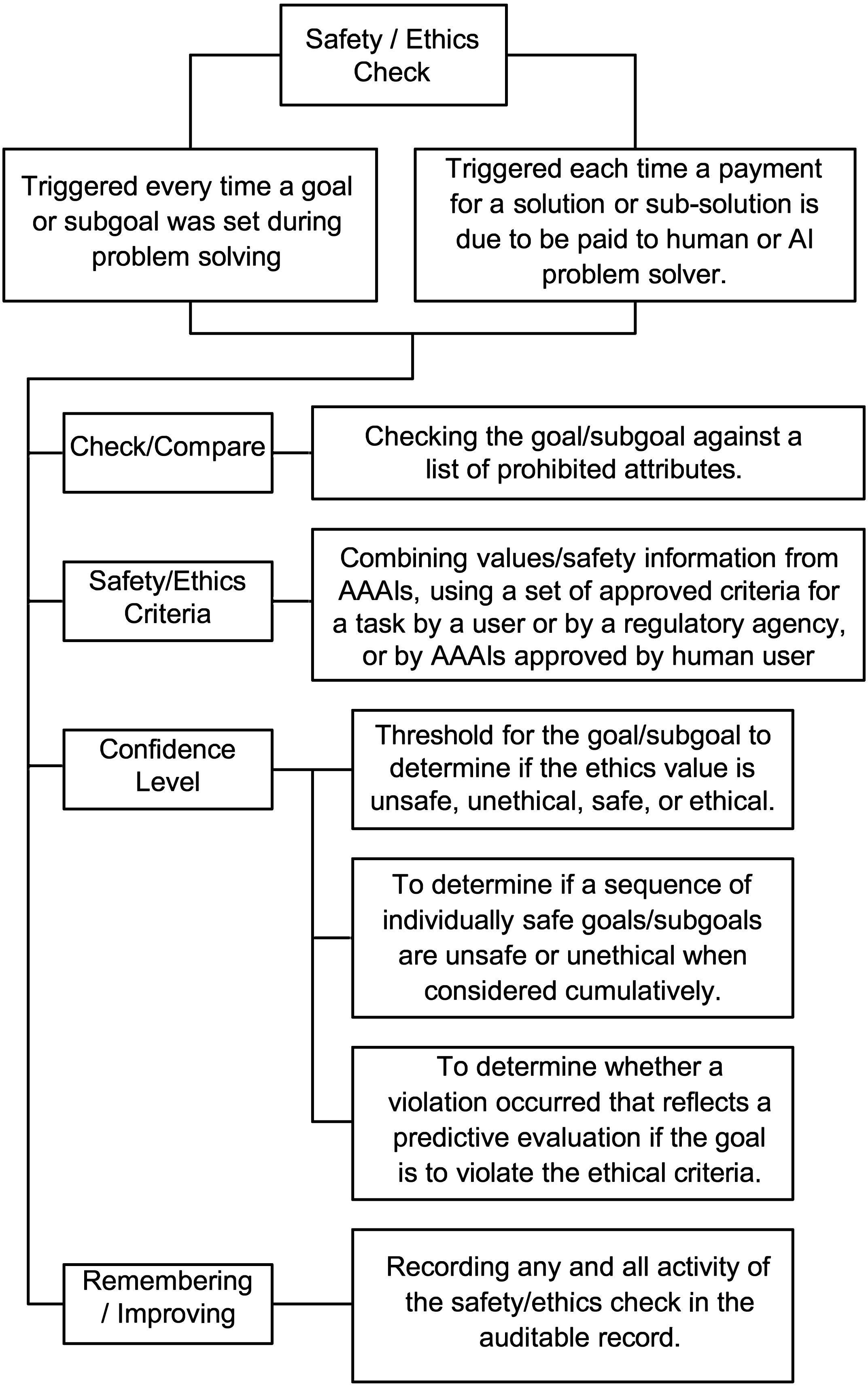

Safety / Ethics Check
Triggered every time a goal or subgoal was set during problem solving
Triggered each time a payment for a solution or sub-solution is due to be paid to human or AI problem solver.
Check/Compare
Checking the goal/subgoal against a list of prohibited attributes.
Safety/Ethics Criteria
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. 19

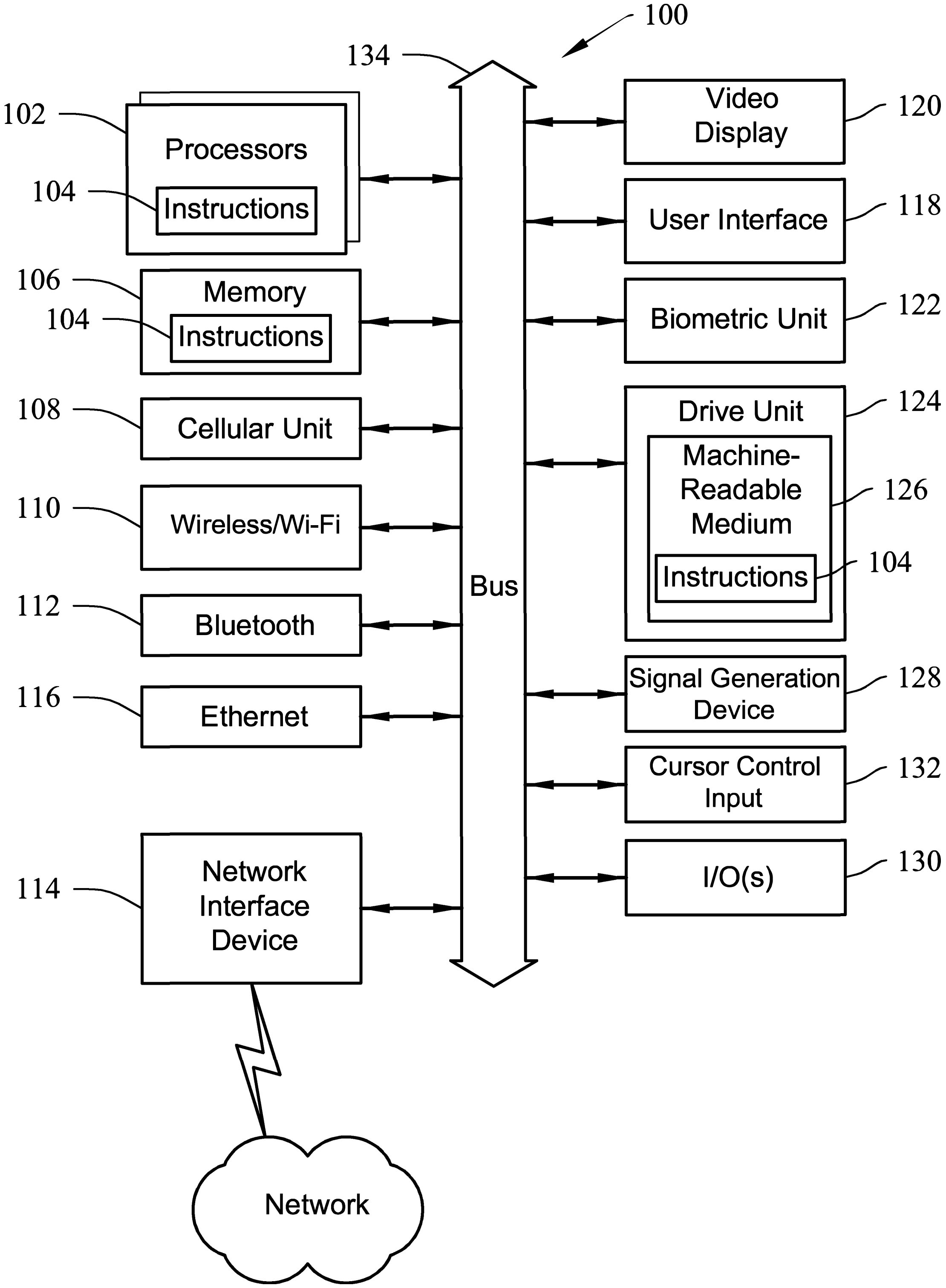

100
134
Bus
102
Processors
104
Instructions
106
Memory
104
Instructions
108
Cellular Unit
110
Wireless/Wi-Fi
112
Bluetooth
116
Ethernet
114
Network Interface Device
Network
Video Display
120
User Interface
118
Biometric Unit
122
Drive Unit
124
Machine-Readable Medium
126
Instructions
104
Signal Generation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

FIG. 20