

SAFE PERSONALIZED SUPER INTELLIGENCE (PSI)

TECHNICAL FIELD

[0001] In some aspects, the present technology relates to a safe personalized super intelligence (PSI) for use in connection with implementing PSI using intelligent agents to develop and continuously improve PSI for each human owner. In some other aspects, the present technology relates to methods associated with creating PSI utilizing a community of multiple Artificial Intelligence (AI) agents or systems and/or multiple pre-customized PSIs all communicating on a collective network, and which all have agreed on a set of agrees on human-aligned values and ethical rules reflecting the values and ethics of the community of PSIs and/or their human users.

[0002] In another aspect, the present technology relates to a PSI that operates on knowledge and values that are pre-trained into the PSI that is at least initially dependent on information provided by or related to the human user.

[0003] In still another aspect, the present technology relates to use of a community of AI agents and/or PSIs to perform safety or ethical checks on other PSIs on the network, to ensure that any task or activity performed by the PSIs follow the agreed upon set of rules.

[0004] In yet other aspects, all activities that are described in this patent disclosure as happening on an external network in which multiple intelligent entities participate in collaborative problem solving, can also be implemented within a single computerized intelligent system where the intelligent entities are all computerized or AI agents or PSIs that reside within that single computerized intelligent system.

BACKGROUND ART

[0005] Previous patent applications describe a path to a development of SuperIntelligent Artificial General Intelligence (AGI)– SuperIntelligent AGI and individual Advanced Autonomous Artificial Intelligences (AAAI). They described specific scenarios involving the existing products and technologies available from several companies. They described how combining data and learning from cross-platform AAAI implementations can accelerate the learning and skills of each AAAI. They described the system and methods for integration in general technical terms and explained how AAAIs can be integrated into an AGI network via a Human-Centered AGI approach.

[0006] The field of AI was named in 1956 at a conference in Dartmouth, MA in the United States that was organized by the computer scientist, John McCarthy. Among the researchers attending the Dartmouth conference were Herbert A. Simon (a future Nobel Laureate) and Allen Newell (a future distinguished computer scientist), both from Carnegie Mellon University.

[0007] Simon and Newell, together with their colleague Cliff Shaw, presented the only working demonstration of AI at the Dartmouth conference. It was a program called the Logic Theorist. The Logic Theorist was an example of the state of early AI efforts where rules defining the behavior of the AI were programmed directly into a computer by human programmers. Interestingly, by programming rules in a general way so as to allow the computer program to pursue goals and subgoals by a variety of means (called “operators”) the Logic Theorist was able to demonstrate creative behavior.

[0008] Specifically, although it was programmed to recreate mathematical proofs from the textbook, Principia Mathematica by Bertrand Russell and Alfred North Whitehead, the Logic Theorist actually found a new proof that was previously unknown both to the programmers of the Logic Theorist and to Russell and Whitehead themselves. Reportedly, Russell and Whitehead were impressed by the Logic Theorist’s new proof and wrote the inventors to say that not only was the Logic Theorist’s proof previously unknown to them, but they wished that they had thought of the proof themselves! Thus, in 1956, at the birth of the field of AI, AI was already capable of creative thought. Of particular relevance to this patent, is the architecture of the Logic Theorist which made use of goals and subgoals – an approach which Newell and Simon subsequently developed further, which was subsequently adopted by many AI systems, and which this patent applies in new and creative ways.

[0009] Research in the field of AI from 1956 to 1986 was primarily dominated by the “expert systems” approach. Humans with programming skills would interview a human expert and represent that expert’s knowledge in a series of programmed rules for the AI. This process was called “knowledge engineering”. The result of the knowledge engineering was an AI program that could behave like a human expert in limited areas. For example, the program MYCIN was developed in the 1970s at Stanford University to act as an expert system in the area of blood infections. E. A. Feigenbaum et al. at Stanford went on to develop an entire series of expert systems in various medical areas in the 1980s. Similar work in expert systems was going on at many other universities as well.

[0010] As more and more expert systems were developed, Newell and Simon looked to the best model of intelligence available – humans – as they strove to improve the performance of AI systems. Their research resulted in a very powerful and broad theory that could describe rigorously how humans solved almost any type of problem. This theory, which elaborated on their earlier work with the Logic Theorist, was known as “search through a problem space”. The theory was described in great detail in their book, Human Problem Solving, published in 1972.

[0011] Dr. Craig Kaplan, the inventor of the AAI patent, studied with Herbert Simon and Allen Newell in the 1980s. He co-authored research with Dr. Simon in the area of creative problem

solving and cognitive science, including publication of an article “Foundations of Cognitive Science” in 1989. Dr. Kaplan realized that the “search through a problem space” architecture proposed by Newell and Simon, could be generalized to enable collective problem solving by millions of humans over the internet. Starting in the late 1990s, Dr. Kaplan began to reduce his ideas to practice in a variety of working systems that actively harnessed the collective intelligence of humans.

[0012] For example, Dr. Kaplan pioneered some of the first practical applications of crowdsourced intelligence around 2000. In 2001, in a presentation at the first Global Brain Conference in Brussels, he outlined his ideas to apply collective intelligence to one of the most difficult and competitive problems in business – beating Wall Street. By 2006, he had designed and implemented the “PredictWallStreet” system that harnessed collective intelligence of millions of humans to get an edge in the stock market. In 2018, that system powered one of the top-ten performing market-neutral hedge funds, thus proving its effectiveness to perform at the highest levels in a complex field competing against some of the smartest humans on the planet.

[0013] In the process of designing and implementing these systems, Dr. Kaplan realized that the “search through a problem space” architecture that worked as a general framework for human problems solving, could be adapted and enhanced to serve as a general architecture for cognition that included both human and AI agents. Further, representing intelligent behavior as a form of problem solving provided a way for many AI agents to interact among themselves, pooling their collective intelligence to create AGI. This “Collective Intelligence” approach, presented here as the AAAI system and method for AGI, represents a faster and more powerful path to AGI compared with existing efforts. Most existing efforts to achieve AGI are primarily focused on training larger Large Language Models (LLMs) using more data, more powerful computers, and better machine learning algorithms. The AAAI approach also has the virtue of enabling humans to participate easily in training and improving the intelligence of AIs, including helping form the AI’s values and ethics – an essential feature to ensure the safe development of AGI.

[0014] While Dr. Kaplan recognized the importance of collective intelligence early on, most other AI researchers became ever-more-focused on a sub-field of AI known as machine learning (ML). Starting in the 1980s, ML began to get traction as a way of getting the AI to learn knowledge on its own, instead of having a knowledge engineer program the knowledge into the AI. However, progress in ML was very slow until a paper showing how to use the “backpropagation of feedback” – one of the first practical reinforcement learning techniques – was published in 1986. After that paper, some AI researchers saw that the future of AI would depend on machines teaching themselves, rather than humans programming them. Unfortunately, the computational and data

requirements for ML were enormous and largely beyond the capabilities of 1980s' or even 1990s' technology.

5 [0015] About three decades of the operation of Moore's law – the doubling of computing power every 18 months or so – were required before the computation ability of technology caught up with what ML algorithms required. During this same time, the amount of data available for training such models, particularly on the internet (which began to take off after 1995 with the advent of web browsers) began to increase.

[0016] An "AI winter," from the 1990s through the first decade of the 2000s, had resulted from overly optimistic ambitions for AI that exceeded the readily available data and computer power. 10 However, by 2010, there was a confluence of abundant computing power, data, and "good enough" ML algorithms. Progress in AI began to accelerate rapidly including the development of improved learning algorithms such as "Transformers".

[0017] As of early 2023, the knowledge engineering approach to creating expert systems has largely been ignored in favor machine learning approaches which have successfully enabled machines to 15 teach themselves how to beat the best human champions at Chess, Go, and any two-player game. Programs like AlphaFold have determined the shapes of millions of proteins in a matter of months whereas the best human experts used to take 4-6 years to accurately determine the shape of a single protein. Natural Language Processing (NLP) – a subfield of AI focused on understanding human language -- has made tremendous progress, resulting in assistants like Amazon's Alexa, Apple's 20 Siri, and most recently Large Language Models (LLMs) like GPT from OpenAI.

[0018] The present technology, scaling, and improvement of LLMs was a watershed moment enabling AI to cross over from being a specialized tool of interest in specific areas (aka "narrow AI") to more general applications. With the release of CHATGPT by OpenAI, the subsequent release of BARD by Google®, the incorporation of GPT into Microsoft's® BING® search engine, 25 and the proliferation of AI companies focused on applying ML approaches widely, a tidal wave of innovation in AI applications is being unleashed. Many individual fields, ranging from medical applications, vehicle navigation, office work, legal work, marketing, sales, education, and even brewing beer are all being revolutionized by application of LLMs, and more broadly, advances in machine learning approaches and capabilities.

30 [0019] However, one goal has remained beyond reach. As of February 28, 2023, except for the present technology detailed in this description, no company or individual has explained how to create a practical system for AGI. The reason: ML alone is not enough to rapidly achieve AGI. Collective Intelligence is also needed.

DISCLOSURE OF TECHNOLOGY

[0020] In view of the foregoing disadvantages inherent in the known approaches to PSI, at least some embodiments of the present technology provide a novel implementation or creation of safe PSI, and overcomes one or more of the mentioned disadvantages and drawbacks of the prior art. As such, the general purpose of at least some embodiments of the present technology, which will be described subsequently in greater detail, is to provide a new and novel safe PSI which has all the advantages of the prior art mentioned herein and many novel features that result in a safe PSI which is not anticipated, rendered obvious, suggested, or even implied by the prior art, either alone or in any combination thereof.

[0021] According to one aspect, the present technology can include system for personalized super intelligence (PSI) using intelligent agents to develop and continuously improve PSI for a human user utilizing a computer system, and additional PSIs, all electronically communicating over a collective network. The system can include a computer system including a processor, a computer-readable storage medium, and program instructions stored on the computer-readable storage medium being executable by the processor to cause the computer system to:

- implement a base-level AI agent on the computer system, wherein the base-level AI agent has already been customized;
- collect media information related to the human user;
- analyze the media information;
- transform the analyzed media information into training data sets;
- differentially weighting the transformed training data sets;
- add knowledge modules to the weighted transformed training data sets;
- locate new sources of data to include to the weighted transformed training data sets;
- apply the weighted transformed training data sets to the base-level AI agent to create a personalized PSI; and
- communicate the personalized PSI with multiple additional PSIs using the network to enable community-based safety features from multiple additional PSIs to the personalized PSI.

[0022] According to another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method comprising the steps of:

- acquiring a base-level AI agent that has previously been customized;
- collecting media information related to the human user;

- analyzing the media information;
transforming the analyzed media information into training data sets;
differentially weighting the transformed training data sets;
adding knowledge modules to the weighted transformed training data sets;
5 locating new sources of data to include to the weighted transformed training data sets;
applying the weighted transformed training data sets to the base-level AI agent to create a user
PSI, wherein the base-level AI agent; and
communicating the user PSI with multiple additional PSIs using the network to enable
community-based safety features from multiple additional PSIs to the user PSI.
- 10 **[0023]** Some embodiments of the present technology can include a step of obtaining, by the human
user, by the base-level AI agent or by the user PSI, any one of or any combination of new training
datasets and training modules to be added to the weighted transformed training data sets.
- [0024]** Some embodiments of the present technology can include a step of monetizing the user PSI
by enabling other human users or other PSIs to access and use the weighted transformed training
15 data sets of the personalized PSI.
- [0025]** In some embodiments, the base-level AI agent can be any one of a pre-trained Large
Language Model (LLM), other tunable and trainable AI agent, and one or more customized AIs
from other human owners.
- [0026]** In some embodiments, the media information can be any one of or any combination of
20 videos of the human user, videos of people and topics related to the human user, photographs or
images of the human user, photographs or images of people and topics related to the human user,
writings related to the human user, journals related to the human user, blogs related to the human
user, posts related to the human user, tweets related to the human user, emails related to the human
user, podcasts related to the human user, recordings related to the human user, auditory content
25 related to the human user, auditory content of people and topics related to the human user, data
collected by third-party vendors that are related to the human user, websites related to the human
user, apps related to the human user, online information related to the human user, social media
information related to the human user, and other AI agents related to the human user.
- [0027]** Some embodiments of the present technology can include a step of providing permission to
30 one or more social media platforms allowing social media content related to the human user to be
accessible by the base-level AI agent or the user PSI.
- [0028]** In some embodiments, the analyzing of the media information can further include the
annotating and categorizing the media information.

[0029] In some embodiments, the analyzing of the media information can utilize one or more algorithms including any one of or any combination of transformer algorithms, deep learning algorithms, transcription algorithms, content and sentiment analysis and summarization, LLMs, and crowdsourced and crowd-supervised human.

- 5 [0030] Some embodiments of the present technology can include a step of compensating, by the human user, one or more human workers on a crowd-sourcing website to review the training data sets, or to suggest refinements to the training data sets.

- [0031] In some embodiments, the step of weighting the transformed training data sets can be performed by adjusting a weight value of any one of the transformed training data sets by the human
10 user utilizing an interface on the computer system that is implementing the base-level AI agent.

[0032] In some embodiments, the step of weighting the transformed training data sets can be performed by any one of or any combination of the additional AI agents, the additional PSIs, and additional human workers each utilizing a computer system.

- [0033] Some embodiments of the present technology can include a step of inputting ethical values to
15 the training data sets by providing a series of interactive dialogs to the human user, the interactive dialogues include predetermined ethical scenarios.

[0034] Some embodiments of the present technology can include a step of mixing the weights of the transformed training data sets to improve the LLM of the base-level AI agent.

- [0035] Some embodiments of the present technology can include a step of authorizing, by the
20 human user utilizing an interface of the base-level AI agent, the LLM to download data and automatically add the data to the training data set or to create a new training data set.

[0036] In some embodiments, the downloaded data can be provided to the human user by way of the computer system for approval by the human user prior to adding to the training data set or to the creation of the new training data set.

- 25 [0037] Some embodiments of the present technology can include a step of cloning the user PSI into one or more cloned user PSIs.

[0038] Some embodiments of the present technology can include a step of training each of the cloned PSIs using training data sets different from each other, wherein the training data sets of each of cloned PSIs each having weights different from each other.

- 30 [0039] Some embodiments of the present technology can include a step of combining the weighted training data sets from the user PSI and one or more of the cloned PSIs to create a combined training data set.

- [0040] Some embodiments of the present technology can include a step of offering the any one of or any combination of the weighted training data sets of any one of or any combination of the cloned PSIs to any one of the additional AI agents and the additional PSIs for use in the training.
- 5 [0041] Some embodiments of the present technology can include a step of training the user PSI with the weighted training data sets of any one of or any combination of the cloned PSIs.
- [0042] Some embodiments of the present technology can include a step of executing, by the human user utilizing the computer system, an automatic process of the user PSI to perform tasks without intervention by the human user unless predetermined parameters triggered.
- 10 [0043] Some embodiments of the present technology can include a step of monitoring, by the base-level AI agent or the user PSI, an activity against a list of prohibited activities and triggering an intervention activity that is provided to the human user.
- [0044] Some embodiments of the present technology can include a step of detecting, by the base-level AI agent or the user PSI, if a gap in the training data sets exists, and if so, then generate interactions with the intelligent entities to obtain data that fills the gaps.
- 15 [0045] In some embodiments, the interactions between more than one of the intelligent entities can be performed in parallel.
- [0046] Some embodiments of the present technology can include a step of providing a task to a plurality of the additional AI agents or the additional PSIs on the network, providing a result to the task from each of the additional AI agents or the additional PSIs, and determining a winning result
- 20 from the results.
- [0047] In some embodiments, the additional AI agents or the additional PSIs can work on the task independently and in parallel with each other.
- [0048] In some embodiments, one of one or more of the additional AI agents or the additional PSIs can be un-customized, and one or more of the additional AI agents or the additional PSIs can be
- 25 customized.
- [0049] Some embodiments of the present technology can include a step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.
- [0050] Some embodiments of the present technology can include a step of utilizing a universal
- 30 problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.
- [0051] Some embodiments of the present technology can include a step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.

[0052] In some embodiments, each of the additional AI agents or the additional PSIs can have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.

5 [0053] In some embodiments, records of activity can be available to any computer device on the network.

[0054] Some embodiments of the present technology can include a step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.

10 [0055] Some embodiments of the present technology can include a step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.

[0056] According to yet another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

15 creating a user PSI by:

acquiring a base-level AI agent that has previously been customized;

collecting media information related to the human user;

analyzing the media information;

transforming the analyzed media information into training data sets;

20 differentially weighting the transformed training data sets; and

applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;

communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional

25 AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics;

recording all actions by the user PSI and the additional AI agents or the additional PSIs in an auditable form on any one of or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs; and

30 monitoring that each of the actions follow the set of rules, and flagging any of the actions that do not follow the set of rules.

[0057] In some embodiments, the recording of the actions can utilize blockchain technology.

[0058] In some embodiments, the recording of the actions can utilize a universal problem solving framework to record any one of or any combination of goals, sub-goals, problem states, and steps taken in a problem solving or cognitive activity that takes place on the network.

[0059] In some embodiments, the recorded actions can be changed only when a majority of the additional AI agents or the additional PSIs on the network provide an authorization for the change.

[0060] Some embodiments of the present technology can include a step of halting any of the flagged actions and applying a preventive operation to prevent a re-occurrence of the flagged action.

5 [0061] Some embodiments of the present technology can include a step of identifying one or more of the additional AI agents or the additional PSIs that provided the flagged action, and controlling a participation of the identified AI agent or PSI on the network.

[0062] Some embodiments of the present technology can include a step of analyzing the flagged action and adjusting any one of or any combination of the training data sets, the set of rules, and
10 attributes of the network based on the analyses of the flagged action.

[0063] According to still yet another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

15 creating a user PSI by:

acquiring a base-level AI agent that has previously been customized;

collecting media information related to the human user;

analyzing the media information;

transforming the analyzed media information into training data sets;

20 differentially weighting the transformed training data sets; and

applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;

providing a task to the user PSI and to a plurality of additional AI agents or additional PSIs on the network;

25 providing a result to the task from each of the user PSI and the additional AI agents or the additional PSIs; and

determining a winning result from the results.

[0064] In some embodiments, each of the additional AI agents or the additional PSIs can work on the task independently and in parallel with each other.

30 [0065] In some embodiments, one of one or more of the additional AI agents or the additional PSIs can be un-customized, and one or more of the additional AI agents or the additional PSIs are customized.

[0066] Some embodiments of the present technology can include a step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.

5 [0067] Some embodiments of the present technology can include a step of utilizing a universal problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.

[0068] Some embodiments of the present technology can include a step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.

10 [0069] In some embodiments, each of the additional AI agents or the additional PSIs can have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.

[0070] In some embodiments, records of activity can be available to any computer device on the network.

15 [0071] Some embodiments of the present technology can include a step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.

[0072] Some embodiments of the present technology can include a step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.

20 [0073] In some embodiments, the network can include the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.

[0074] According to another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

25 a) creating a user PSI by:
acquiring a base-level AI agent that has previously been customized;
collecting media information related to the human user;
analyzing the media information;
30 transforming the analyzed media information into training data sets;
differentially weighting the transformed training data sets; and
applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;

- b) communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics;
- c) performing a baseline performance of the user PSI and the additional AI agents or the additional PSIs utilizing a variety of standardized tasks;
- d) creating one or more versions of the user PSI each different from each other;
- e) determining a performance of the network of the user PSI, the versions of the user PSI, and the additional AI agents or the additional PSIs;
- f) assigning a credit value or a blame value to one or more elements that resulted in superior or inferior performance of the baseline performance; and
- g) determining which of the elements are superior elements that resulted in an increase in performance and retaining the superior elements for future use.
- [0075]** Some embodiments of the present technology can include a step of determining which of the elements are inferior elements that resulted in a decrease in performance and change an attribute of the inferior elements.
- [0076]** Some embodiments of the present technology can include a step of repeating step c) utilizing the superior elements and the changed inferior elements.
- [0077]** Some embodiments of the present technology can include a step of repeating steps c)-g) until no more superior elements are detected.
- [0078]** In some embodiments, the network can include the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.
- [0079]** Some embodiments of the present technology can include a step of recording all actions by the user PSI and the additional AI agents or the additional PSIs in an auditable form on any one of or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs.
- [0080]** In some embodiments, the standardized tasks can be any one of or any combination of problem solving tasks using a shared universal problem solving architecture, ethical and safety scenarios designed to determine whether any one of or any combination of the use PSI, and the additional AI agents or the additional PSIs behaves safely and ethically, standardized intelligence tests and assessments that have been designed to test an intelligence of the intelligent entities, tasks configured to measure a degree of “hallucination” or production of erroneous results, tasks that come from a variety of domains of expertise, tasks that incorporate different cultural or group norms for behavior, tasks that are random permutations of existing standardized tasks, and tasks that are dynamically created based on changing conditions.

[0081] In some embodiments, the versions of the user PSI can be created by any one of or any combination of: human adjustment of parameters, weights or data encoding a knowledge of the user PSI; autonomous adjustment of the parameters, weights or data encoding the knowledge the user PSI; random variation of the parameters, weights or data encoding the knowledge of the user PSI; 5
subjecting the user PSI to different training regimes or amounts of training; creating the different versions of the user PSI sequentially, and creating the different versions of the user PSI in parallel.

[0082] In some embodiments, the versions of the user PSI can be created by directly combining the parameters, weights or data encoding the knowledge of multiple PSIs by computing average weight values, with weighting more recently created or more complex versions of the user PSI more than 10
older or less complex versions of the user PSI.

[0083] In some embodiments, the versions of the user PSI can be created by random or deliberate methods for estimating which of the versions are to have beneficial results, wherein the methods for estimating being any one of or any combination of comparing a degree of match between a knowledge of one of the versions of the user PSI and a statistical frequency of tasks that are 15
submitted to the network, comparing an overlap in knowledge of the additional PSIs that are different on the network such that changes are made to optimize a performance of groups of the versions of the user PSI given a characterization of an entire network knowledge compared to a characteristics of a problem that the network is expected to solve, and using hill climbing or gradient descent for optimizing parameters of any one of the versions of the user PSI.

[0084] According to another aspect, the present technology can include a method for PSI utilizing a single computerized intelligent system including multiple AI agents residing in the single computerized intelligent system. The method can include:

- acquiring a base-level AI agent that has previously been customized, the base-level AI agent resides in a single computerized intelligent system;
 - 25 collecting media information related to a human user associated with the base-level AI agent; transforming the analyzed media information into training data sets; differentially weighting the transformed training data sets; applying the weighted transformed training data sets to the base-level AI agent to create a user PSI; and
 - 30 communicating the user PSI with multiple additional PSIs to enable community-based safety features from the additional PSIs to the user PSI.
- training a base Large Language Model (LLM) of an AI agent with guardrails including attributes associated with any one of or any combination safety, ethics and knowledge, the AI agent residing in a single computerized intelligent system;

- customizing the base LLM to an ethics profile;
combining ethical information from multiple additional AI agents residing in the single
computerized intelligent system, the additional AI agents being different to that of the AI
agent, the additional AI agents residing in the single computerized intelligent system;
5 refining a set of values of the base LLM based on problem solving of a problem request; and
updating the base LLM with the combined ethical information and the refined set of values
thereby allowing for a scalable AGI.

[0085] There has thus been outlined, rather broadly, features of the present technology in order that
the detailed description thereof that follows may be better understood and in order that the present
10 contribution to the art may be better appreciated.

[0086] Numerous objects, features and advantages of the present technology will be readily
apparent to those of ordinary skill in the art upon a reading of the following detailed description of
the present technology, but nonetheless illustrative, embodiments of the present technology when
taken in conjunction with the accompanying drawings.

15 [0087] As such, those skilled in the art will appreciate that the conception, upon which this
disclosure is based, may readily be utilized as a basis for the designing of other structures, methods
and systems for carrying out the several purposes of the present technology.

[0088] It is therefore an object of the present technology to provide a new and novel safe PSI that
has all of the advantages of the prior art AGI systems and methods and none of the disadvantages.

20 [0089] It is another object of the present technology to provide a new and novel safe PSI that may
be easily and efficiently implemented and marketed.

[0090] An even further object of the present technology is to provide a new and novel safe PSI that
has a low cost of implementation with regard to both resources and labor, and which accordingly is
then susceptible of low prices of sale to the consuming public, thereby making such safe PSI
25 economically available to the buying public.

[0091] Still another object of the present technology is to provide a new safe PSI that provides in the
system and methods of the prior art some of the advantages thereof, while simultaneously
overcoming some of the disadvantages normally associated therewith.

30 [0092] For a better understanding of the present technology, its operating advantages and the
specific objects attained by its uses, reference should be made to the accompanying drawings and
descriptive matter in which there are illustrated embodiments of the present technology. Whilst
multiple objects of the present technology have been identified herein, it will be understood that the
following description is not limited to meeting most or all of the objects identified and that some
embodiments of the present technology may meet only one such object or none at all.

BRIEF DESCRIPTION OF THE DRAWINGS

[0093] The technology will be better understood and objects other than those set forth above will become apparent when consideration is given to the following detailed description thereof. Such description makes reference to the annexed drawings wherein:

[0094] FIG. 1 is a flow chart illustrating an embodiment of the subsystems utilizable in the AAAI system and method of the present technology.

[0095] FIG. 2 is a block diagram illustrating an exemplary process of the overall process utilizable with the present technology.

[0096] FIG. 3 is a flow chart illustrating an exemplary embodiment of the system and methods for creating a scalable ethical and safe AGI or PSI from the collective intelligence of AAAs and humans utilizable with the present technology.

[0097] FIG. 4 is a flow chart illustrating an exemplary embodiment of the scalable universal problem solving system and methods for human-centered AGI, with relevance for PSIs, constructed in accordance with the principles of the present technology.

[0098] FIG. 5 is a flow chart illustrating an exemplary embodiment of the scalable solution learning subsystem or process.

[0099] FIG. 6 is a flow chart illustrating an exemplary embodiment of the scalable natural language to problem solving language translator subsystem or process.

[00100] FIG. 7 is a flow chart illustrating an exemplary embodiment of the scalable reputational component subsystem or process for the human, AI, and/or PSI problem solving agents.

[00101] FIG. 8 is a flow chart illustrating an exemplary embodiment of the scalable safety and ethics checks subsystem or process, wherein AIs might also be PSIs.

[00102] FIG. 9 is a diagram illustrating features and functions of the Problem Solving architecture including the Tree structure used by the scalable WorldThink protocol, wherein AAAs might also be PSIs.

[00103] FIG. 10 is a diagram illustrating various use cases for domain-specific problems which depend upon the underlying WorldThink protocol, and which together help form the basis for an AGI system capable of solving a wide range of problems, wherein the AAAs identified in the Figure might also be PSIs.

[00104] FIG. 11 is a diagram illustrating some of the steps in the universal problem solving framework that is part of the WorldThink protocol and used by the AAAI system and which also can be used by PSIs.

[00105] FIG. 12 is a flow chart illustrating some of the basic problem solving functionality

supported by the WorldThink protocol utilizable with the AAAI system and method of the present technology, and wherein solvers might be PSIs.

[00106] FIG. 13 is a flow chart illustrating some of the basic problem solving functionality supported by the WorldThink protocol utilizing two problem solvers, which could be PSIs, collaborating to solve a client problem.

[00107] FIG. 14 is a flow chart illustrating an exemplary customization process of an AAAI system, wherein the AIs in the Figure might also be PSIs.

[00108] FIG. 15 is a flow chart illustrating an exemplary problem solving process utilizing a common cognitive architecture implemented in an AI system, wherein the AIs in the Figure might also be PSIs.

[00109] FIG. 16 is a flow chart illustrating an exemplary problem solving process utilizing a common cognitive architecture implemented in a collective network of AI systems wherein the AIs in the Figure might also be PSIs.

[00110] FIG. 17 is a flow chart illustrating an exemplary implementation of the PSI of the present technology.

[00111] FIG. 18 is a flow chart illustrating an exemplary implementation of leveraging the abilities of the PSI of the present technology.

[00112] FIG. 19 is a flow chart illustrating an exemplary implementation of the community-based safety mechanism of the present technology.

[00113] FIG. 20 is a flow chart illustrating an exemplary implementation of the recording of all actions by the PSIs on the network.

[00114] FIG. 21 is a flow chart illustrating an exemplary implementation of checking of cognitive activity on the network.

[00115] FIG. 22 is a flow chart illustrating an exemplary implementation of the competitive evolution of intelligence and/or performance of the PSIs.

[00116] FIG. 23 is a flow chart illustrating an exemplary implementation of joining multiple PSIs on a network of PSIs with agreed upon rules and methods for interaction.

[00117] FIG. 24 is a flow chart illustrating an exemplary implementation of baseline performance of individual PSIs, groups of PSIs, and/or the entire network of PSIs, on a variety of standardized tasks.

[00118] FIG. 25 is a flow chart illustrating an exemplary implementation of producing different versions of the individual PSIs and/or different combinations of PSIs.

[00119] FIG. 26 is a schematic block diagram illustrating an exemplary electronic computing device that may be used to implement an embodiment of the present technology.

[00120] The same reference numerals refer to the same parts throughout the various figures.

DETAILED DESCRIPTION OF THE TECHNOLOGY

5 Definitions

[00121] Artificial Intelligence (AI) – A non-human entity capable of behavior that most humans would consider intelligent in at least one area, or in some respect.

[00122] Artificial General Intelligence (AGI) – Conventionally refers to an AI that is capable of doing all (or almost all) intellectual tasks that an average human could do. However, it should be
10 clear that any AGI capable of learning and self-improving will not remain at the AGI level very long but will rapidly progress to becoming SuperIntelligent AGI that can do all intellectual task as well or better than the average human. So, for purposes of this description, “AGI” will refer to either a conventional AGI system or a “SuperIntelligent” AGI. In this description, the AGI is described as being implemented by a system and associated methods.

15 [00123] Advanced Autonomous Artificial Intelligence (AAAI) – An AI capable of independent or semi-independent (supervised) intelligent action. An AI agent. An individual AAAI can be specified, customized, and put into useful action via the systems and methods of this AAAI present technology. A group of AAAIs can cooperate and combine their intelligence to create an integrated AGI system. A sufficiently advanced AI agent can also act as an AGI system which may include
20 other less advanced AI agents within itself.

[00124] AAAI.com – A platform, company, website, and/or project that implements this the present technology and supports the development, customization, and use of AAAI agents and the AGI that results from the combined action, knowledge, or intelligence of multiple AAAIs, via collective intelligence of AAAIs and/or humans, as specified in this and related technologies.

25 [00125] AI Ethics – The ethics adopted by an AI or AGI that describe what is right and wrong in given contexts.

[00126] Alignment Problem – The problem that arises when AI Ethics are not aligned with Human Ethics resulting in AI or AGI taking actions that humans consider unethical and/or which are dangerous to individual humans or the human race.

30 [00127] Base AI – An AI, AI Agent, AAAI, SLM or LLM that has been trained generally but has not yet been customized with information from individual users or with information for specific tasks.

[00128] Collective Intelligence (CI) – The intelligence that emerges when multiple intelligent entities are focused on solving a common problem, or when the knowledge from multiple intelligent

entities is pooled to overcome limits of bounded rationality. Collective Intelligence historically has been human collective intelligence, but AGI is based on collective intelligence of both human and AI agents (including PSIs) and can also result from multiple AAAIs with or without human participation in the system. Active CI results from intelligent entities (e.g., humans or machines) taking steps that are useful in solving a problem or participating actively in other intellectual endeavors. For example, when multiple humans explicitly tell an advertiser what type of ads they want to see, the humans are exhibiting active CI. Passive CI results from analyzing the behavior of an intelligent entity (e.g., a human or a machine) even if such behavior was not directly related to solving the problem for which the analysis is used. For example, when an AI or other system analyzes which web pages a (group of) human(s) visit on the web, and then uses that analysis to direct targeted ads to the human(s).

[00129] Ethics/Values (“Ethics”) – A subset of knowledge that provides a sense of purpose to an intelligent entity and that serves to constrain allowable actions or operations based on what is asserted to be “right” or “wrong” behavior in a given context. Specifically, Ethics should be considered premises from which an intelligent entity can reason or logically compute the best course of action to achieve the goals or intents consistent with the ethical premise. Just as premises must be accepted “as given” in systems of logic, so too, fundamental ethics or ideas of what is right and what is wrong must be accepted as premises, from which starting point an intelligent entity can propose rational actions to realize those values or ethics.

[00130] Hallucination/Artificial Hallucination – A phenomenon wherein a large language model (LLM), often a generative AI chatbot or computer vision tool, perceives patterns or objects that are nonexistent or imperceptible to human observers, or creates outputs that are nonsensical, inaccurate, misleading or false.

[00131] Human Ethics – The ethics asserted by human beings which describe what is right and wrong in given contexts.

[00132] Intelligent Entities or Entity – A human utilizing a computer system, an AI agent or system, a clone of an AI agent or system, an AAAI agent or system, and/or a clone of an AAAI agent or system, which participates in providing a problem, a subproblem, a goal and/or a subgoal, and/or participates in any problem solving activity on a problem, a subproblem, a goal and/or a subgoal. In the case of multiple intelligent entities within a single computer system, intelligent entities also refers to the sub-programs of parts of that overall computer program that function as an intelligent entity within the larger collection of simulated or programmed entities. PSIs are also intelligent entities.

[00133] Large Language Model (LLM) – A type of AI that can accept natural language as an input and generate natural language as an output. Typically, LLMs are trained using ML techniques on large datasets so that they can emulate intelligent conversation or other forms of interaction with humans in natural language. Variants of LLMs can also be trained to take language as input and generate images or visual representations as output; or they can take images and visual representations as input and generate language and/or image and/or visual representations as output. For the purposes of this patent, we will refer to all such systems as LLMs even though the image-based models do not always need to accept text as the input or the output. LLMs can also act as a type of AI agent and are sometimes referred to as such in the present technology. For purpose of this disclosure, Small Language Models (SLMs) are also included in the definition of LLM.

[00134] Machine Learning (ML) – A sub-field that is concerned with developing AI by enabling machines to teach themselves or learn their knowledge rather than such knowledge being explicitly programmed into them (as would be the case with an Expert System AI developed via classical knowledge engineering methods).

[00135] Narrow AI – An AI that performs at human or at super-human levels in a relatively restricted domain such as game playing, brewing beer, analyzing legal contracts, etc. Narrow AI is contrasted with AGI that can perform at human level at ALL intellectual tasks. Some AIs are narrower than others, for example driving a car requires more general ability than playing chess but not as much as an AGI would have.

[00136] Prohibited Attributes – Requests, goals, problems, terms, phrases, questions, answers, solutions, information and the like that are determined or set as being illegal, immoral, unethical, dangerous, deadly and the like. For example, requesting information for getting Molotov Cocktails through airport security.

[00137] Safety – Generally, the concern for human safety and survival is distinct from ethics and values.

[00138] Safety Feature – An aspect of the design or operation of the present technology which increases the safety of one or more humans, often by helping increase the probability that AI ethics align with human ethics, thus surmounting the Alignment Problem.

[00139] Training/Tuning/Customization – Conventionally the term “training” is used to denote training a network (e.g., LLM) to behave intelligently. Tuning refers to activities that fine-tune the trained base model so that it performs even better, typically at specific tasks. Customizing refers to a wide variety of activities including, but not limited to, training and tuning that make an AI uniquely suited for the purposes of a given user(s) or application(s). For purposes of this description, Training, Tuning, and Customization are used interchangeably with the understanding that although

techniques vary, and the degree and type of effort involved varies, the aim of all three is to adapt the AI and make it behave more intelligently or more uniquely suited to a particular user(s) or application(s).

[00140] Weights/Weights of the Network – In the field of machine learning, many systems learn by adjusting the weights in a neural network architecture that can be represented as a network of nodes and links between nodes. The weight of a link connecting two nodes, for example, may correspond to the strength of association or connection between the whatever nodes represent. These weights can also represent excitatory or inhibitory connections between concepts, as in a neural network representation. The learning of an entire AI system, such as a LLM or more generally any AI agent that has learned via back-propagation of error, transformer algorithms or any of the machine learning methods for establishing and modifying strengths of connections between nodes (also called “parameters” in some models) can be represented as a matrix of numbers corresponding to the weights between the nodes in the network. Weights / Weights of the Network in this description refer to this numerical information, often but not necessarily stored in a matrix or vector representation. By combining, manipulating, or otherwise changing this numerical information, the learning, knowledge, or expertise and behavior of the system can be changed.

Overview of the Present Technology

[00141] The present technology is comprised of systems and methods for implementing a Personalized SuperIntelligence (“PSI”), superior to all currently existing forms of AI both in scope and intelligence. Since this PSI is self-improving, it will become exponentially more intelligent than the owner who creates it. However, due to the unique method of creation described in this patent, the PSI will be entirely safe and dedicated to the service of the owner. Safe SuperIntelligence by design is the essence of this patent. The present technology is unlike any AI system previously created, and has intelligence levels, safety, and useful benefits far beyond the current state-of-the-art AI assistants.

[00142] Think of your PSI as your personal, super-smart, AI agent. It does your bidding and learns about you and what you want over time. It is far more effective and efficient than you are at most everyday online tasks. It offers excellent advice. And best of all it knows you, can relate to you, and thinks like you. If you want to expand the capabilities of your PSI, it is as simple as trading data with a friend, or buying or selling data on an online marketplace, and then mixing that data into your PSI such that it provides additional intelligence, skills, and knowledge. Over time, your PSI can acquire the wisdom of billions of humans and AI agents. And after acquiring that knowledge, it can

run simulations based on your goals and purposes to improve further and act on your behalf (with your permission) as you see fit.

[00143] Your material wealth can be multiplied by your PSI, acting as your investment advisor and agent. Your time can be freed up for spiritual, artistic or other pursuits.

5 [00144] Your PSI can be cloned and leased out. Its data and wisdom can be packaged and sold. A thousand versions of it can handle a thousand different errands and missions for you simultaneously, under the direction of other cloned PSIs that are also under your direction and serving your interests.

[00145] These are some of the benefits that the present technology unleashes for an individual owner. For society at large, multiple PSIs, pooling their intelligence and participating in a PSI
10 network, can serve as a Planetary Intelligence, providing benefits for all people and our planet.

[00146] While the above-described devices fulfill their respective, particular objectives and requirements, the aforementioned devices or systems do not describe a safe PSI that allows implementing PSI using intelligent agents to develop and continuously improve PSI for each human
15 owner. The present technology additionally overcomes one or more of the disadvantages associated with the prior art.

[00147] A need exists for a new and novel safe PSI that can be used for implementing PSI using intelligent agents to develop and continuously improve PSI for each human owner. In this regard, the present technology substantially fulfills this need. In this respect, the safe PSI according to the present technology substantially departs from the conventional concepts and designs of the prior art,
20 and in doing so provides an apparatus primarily developed for the purpose of implementing PSI using intelligent agents to develop and continuously improve PSI for each human owner.

[00148] In the following description, for purposes of explanation and not limitation, specific details are set forth, such as particular embodiments, procedures, techniques, etc. in order to provide a thorough understanding of the present technology. However, it will be apparent to one skilled in the
25 art that the present technology may be practiced in other embodiments that depart from these specific details.

[00149] It can be appreciated that the present technology provides a technical effect, contribution and solution with a technical implementation of multiple customized AAAI systems communicating over a collective intelligence neural network, in combination with all the AAAI systems each
30 utilizing a common cognitive architecture including one or more problem solving protocols for generating one or more solutions or answers to a problem request, and providing the solutions or answers to a user for approval. Where the customization of the AI system resulting in the AAAI includes input from human users for training the AI, the AAAI, or the PSI. Further technical contribution or solution can be where the multiple customized AAAI or PSI systems can include one

or more cloned AAAs or PSIs that can each be customized independently of a parent AAAI or PSI and independent of other cloned AAAs or cloned PSIs of the same system.

5 [00150] Still another technical contribution and solution is for the creation and/or implementation of a PSI that is self-improving and that will become exponentially more intelligent while being safe and ethical. This technical contribution can be further provided in the creation of the PSI by acquiring a base-level AI agent that has previously been customized, and utilizing differentially weighted media information related to the human use in the training of the PSI, in combination with training information acquired by a community of additional AIs and/or PSIs all communicating on a collective network..

10 [00151] Still yet another technical contribution and solution is in that all the AIs and PSIs communicate on the collective network enable community-based safety features and/or learning, with all the AIs and PSIs agreeing on a set of rules governing safety and ethical that are human-aligned values to prevent any harmful actions by any malevolent PSIs.

15 [00152] It can be appreciated that the present technology is found outside of computer program exclusion and/or abstract idea interpretation. This can in part be found in the technical contributions and solutions provided by the present technology, the utilization of specific training input that is external to a computer, and the providing of the solution or answer external to a computer.

20 [00153] One reason AGI has been so elusive is that specific knowledge and expertise from diverse fields must be creatively combined in an invention to achieve AGI. Another reason the development of AGI has been non-obvious, is that almost all AI researchers are focused on trying to improve existing narrow AI systems via ever more complex and extensive machine learning approaches.

25 [00154] The fact that AGI has resisted attempts by thousands of others -- despite the expenditures of huge sums of money -- and the fact that specialized knowledge in relatively obscure fields had to be combined with mainstream AI approaches in the present technology, argue strongly for the novelty and creativeness of the present technology.

[00155] The present technology describes the system and methods not only to achieve AGI, but also to achieve it rapidly, and most importantly, safely.

30 [00156] It is possible to influence the evolution of AGI in a positive direction. The best way we can do this is by adopting the safest possible path to the development of AGI and ensuring that humanity follows that path. In turn, the best way to ensure that humanity follows the safest path, is to show that the safest path to AGI is also the fastest and therefore most desirable path to AGI. These considerations, the desire to illuminate the fastest path, which is also the safest path, is motivation for the development of the present technology.

[00157] While the above-described devices fulfill their respective, particular objectives and requirements, the aforementioned devices or systems do not describe a system and methods for safe, scalable, artificial general intelligence that allows scaling by using a combination of human users and multiple AI systems to train other AI systems by combining values and ethical knowledge of the human users and the multiple AI systems for training. The present technology additionally overcomes one or more of the disadvantages associated with the prior art.

[00158] A need exists for a new and novel system and methods for safe, scalable, artificial general intelligence that can be used for scaling by using a combination of human users and multiple AI systems to train other AI systems by combining values and ethical knowledge of the human users and the multiple AI systems for training. In this regard, the present technology substantially fulfills this need. In this respect, the system and methods for safe, scalable, artificial general intelligence according to the present technology substantially departs from the conventional concepts and designs of the prior art, and in doing so provides an apparatus primarily developed for the purpose of scaling by using a combination of human users and multiple AI systems (including without limitation, PSIs) to train other AI systems by combining values and ethical knowledge of the human users and the multiple AI systems for training.

[00159] In the following description, for purposes of explanation and not limitation, specific details are set forth, such as particular embodiments, procedures, techniques, etc. in order to provide a thorough understanding of the present technology. However, it will be apparent to one skilled in the art that the present technology may be practiced in other embodiments that depart from these specific details.

[00160] In the following description, for purposes of explanation and not limitation, specific details are set forth, such as particular embodiments, procedures, techniques, etc. in order to provide a thorough understanding of the present technology. However, it will be apparent to one skilled in the art that the present technology may be practiced in other embodiments that depart from these specific details.

[00161] It can be appreciated that the present technology provides a technical effect, contribution and solution with a technical implementation of multiple customized AAAI (or PSI) systems communicating over a collective intelligence neural network, in combination with all the AAAI (or PSI) systems each utilizing a common cognitive architecture including one or more problem solving protocols for generating one or more solutions or answers to a problem request, and providing the solutions or answers to a user for approval. Where the customization of the AI system resulting in the AAAI (or PSI) includes input from human users for training the AI or the AAAI (or PSI). Further technical contribution or solution can be where the multiple customized AAAI systems can

include one or more cloned AAAs (or PSIs) that can each be customized independently of a parent AAA (or PSI) and independent of other cloned AAAs (or PSIs) of the same system.

5 [00162] Still another technical contribution and solution is for the faster and safer creating of scalable AGI that utilizes human input in training and customization for imparting human ethical attributes to the AAA, PSI, and/or AGI.

[00163] Still yet another technical contribution and solution is for scalably train AI systems and/or agents with a combination of safety and ethical information from many individual AI agents (or PSIs) to achieve a representative and statistically valid sample of human ethics and values covering a wide range of scenarios. A further technical contribution can be found in that the present
10 technology includes methods for combining the information from many agents and assembling optimal combinations of such agents for providing scalable training of AI, PSI, or AGI.

[00164] It can be appreciated that the present technology is found outside of computer program exclusion and/or abstract idea interpretation. This can in part be found in the technical contributions and solutions provided by the present technology, the utilization of specific training input that is
15 external to a computer, and the providing of the solution or answer external to a computer.

[00165] The AAA approach to developing safe AGI is fundamentally a Collective Intelligence (CI) approach. The source of intelligence is not a monolithic LLM, SLM or super-advanced AI, but rather a collection of intelligent agents which can be both human and AI. Component sub-tasks in developing AGI include, without limitation, training individual AI agents or PSIs, combining
20 knowledge (including without limitation subjective values and ethical knowledge) from different agents effectively and efficiently, scaling the AGI, and continuously improving/updating the AGI.

[00166] Current approaches – such as Reinforcement Learning with Human Feedback (RLHF) and Constitutional Learning – are failing to effectively and scalably train AI to be ethical and safe. The present technology describes a scalable system and methods that are superior to current approaches.
25 In one aspect, the present technology can include the combination of safety and ethical information from many individual AI agents to achieve a representative and statistically valid sample of human ethics and values covering a wide range of scenarios. The present technology can include methods for efficiently covering a wide range of ethical situations and dynamically addressing new situations as they emerge. Methods for combining the information from many agents and assembling optimal
30 combinations of such agents are also presented. These methods can be used not only to improve safety using ethical knowledge but also to create superintelligent systems that combine many other types of knowledge. Safe AGI and SuperIntelligence can be achieved via the collective intelligence approach described in this description of the present technology. A detailed scenario, using the

company META® as an example, illustrates one exemplary implementation of the present technology.

[00167] Methods for dynamically updating knowledge are also presented. Successful implementation of the present technology will increase the chances that AI, PSIs, AGI, and SuperIntelligence remain aligned with human values even when such systems greatly exceed humans in intelligence.

[00168] Advanced Autonomous Artificial Intelligence (AAAI) is a set of systems and methods for developing Artificial General Intelligence and SuperIntelligent Artificial General Intelligence (collectively “AGI”) in a rapid and safe manner for the benefit of humankind. In contrast to other approaches to the development of AGI, the AAAI present technology achieves a faster and safer path to AGI by relying, at least initially, on the involvement of (ideally many millions of) humans minds in the AGI training, operation, and safety/supervisory functions.

[00169] The AAAI present technology can achieve AGI by enabling users to first customize and clone their own AIs or PSIs. These customized AIs (AAAI), and/or PSIs, participate in problem solving and other intellectual activities on a network consisting of other AAAs, PSIs, and humans. Although each AAAI (or PSI) on its own may lack the breadth of skills and knowledge to be an AGI, collectively the AAAs (initially with help from humans on the network) form an AGI that will quickly surpass average human ability in all intellectual endeavors.

[00170] Some aspects of the present technology can include: 1) the system and methods to customize AIs with the unique knowledge, skills, and ethical values of the users; 2) the universal problem solving architecture that allows AAAs to interact productively with each other and with humans on intellectual tasks; 3) the network where the interactions takes place; 4) the methods for integrating the knowledge and ethics of individual AAAs into an AGI; and 5) the methods for learning and continuous improvement so that the AAAs and the AGI become smarter and more ethical over time. Involvement of humans as customizers of their AAAs and participants on the network is an essential feature of the present technology which not only accelerates the development of AGI, but also makes AGI safer by providing a mechanism for the ethical values of millions of humans to be adopted by and reflected in the AGI.

[00171] One implementation of the AAAI system of the present technology has a focus on safety and is implemented via five sub-systems and associated methods, as illustrated in FIG. 1. The five sub-systems of the AAAI system are: 1) AAAI Customization, 2) AAAI Architecture, 3) AAAI Network, 4) AAAI Integration, 5) AAAI Improvement. The acronym SCAN--II (Safe, Customizable, Architecture and Network -- Integrated and Improving) describes the present technology in the exemplary implementation. Other combinations of subsystems, and variations of

each subsystem, are also possible. Safety features have been designed into each sub-system in an effort to provide redundant safety checks in the event one or more sub-systems are omitted from a particular implementation.

[00172] The five sub-systems of the AAAI system can be further described as:

- 5 1) A base level Large Language Model (LLM), Small Language Model (SML), or other AI system can be customized to reflect the knowledge of an individual, group of individuals, or organization and designated an Advanced Autonomous Artificial Intelligence (AAAI).
- 2) The customized AAAI can be enabled to participate in problem solving using a universal problem solving architecture that is compatible with both human and AI agents.
- 10 3) The problem solving-enabled AAAI participates in problem solving activity, including but not limited to:
 - planning, problem solving, and other types of sequential, multi-step cognitive activity, on a network of intelligent agents;
 - generate and select operators that reduce a difference between a current state of problem solving
 - 15 and a desired state based on the goal/subgoal;
 - setting of a subgoal towards achieving the goal;
 - utilizing hierarchy until an actionable goal is set that can be acted on by the operator; and
 - analyzing the auditable record to determine recommendations for improvement of the problem solving process to achieve a solution to the goal/subgoal.
- 20 4) Multiple AAAIs, or PSIs, on the network can be integrated to achieve AGI; or AI capable of intelligent (or super-human level) behavior across a wide range of tasks.
- 5) The individual AAAIs, the problem solving network, and/or the integrated system of multiple AAAIs continuously improve via a variety of means, including but not limited to, redirecting the efforts of individual AAAIs and/or the integrated AGI towards the task of improving the system
- 25 and/or components of the system.

[00173] The sub-systems or new sub-systems can include any one of or any combination of:

- 1) Safety/ethics check - Comparing a goal or subgoal against a list of prohibited attributes and assigning an ethics value based on a result of the comparison. Checking the goal/subgoal against a list of prohibited attributes. Combining values/safety information from AAAIs, using a set of
- 30 approved criteria for a task by a user or by a regulatory agency or by AAAIs approved by human user. Establishing or using a threshold for the goal/subgoal to determine if the ethics value is unsafe, unethical, safe, or ethical. Determining if a sequence of individually safe goals/subgoals are unsafe or unethical when considered cumulatively. Determining whether a violation

occurred reflects a predictive evaluation if the goal is to violate the ethical criteria. Recording any and all activity of the safety/ethics check in the auditable record.

- 2) AAAI matching - Detecting and identifying additional AAAIs, or PSIs, that each have a criteria related to one or more goal or subgoal criteria.
- 5 3) Remembering and/or improving – Recording activity, comparing with successful or unsuccessful progress towards the problem solutions, determining which activity to keep active or forget.
- 4) AAAI learning - Learning including a procedural learning process that utilizes information provided by intelligent entities such as human users equipped with computers, AAAIs, or PSIs.
- 10 Recording activity, comparing with successful or unsuccessful progress towards the problem solutions, determining which activity to keep active or forget. Assigning credit value or blame value to a group of content of the problem solving activity. A set of prompts provided to the user and information received based on the prompts. Updating AAAIs with the group of content determined as active. The group of content can be, but not limited to, a set of prompts provided
- 15 to the user and information received based on the prompts, all of which being recorded in the auditable record. Optionally, the problem solving activities can include the group of content.

EXAMPLE USER SCENARIOS

[00174] It may be helpful to describe some user scenarios that provide a sense of how the present technology can operate in some of the aspect implementations. An exemplary process is illustrated in FIG. 2.

[00175] In one aspect, a user “visits” AAAI.com via the user’s computer, cell phone, PDA, or goggles. AAAI.com would interact with the user via a web-based interface, a phone app, custom software for the PDA, or a metaverse / virtual reality environment. The mode of interaction could be physical via a keyboard, mouse, or gestural interface; voice-based via a microphone input coupled to natural language understanding and generation systems; or video-based as in the case where the user becomes an avatar in a virtual reality setting or in the metaverse.

[00176] The initial interaction would include setting up the user’s account, which might be free or paid. This would involve an account name and password or other authentication mechanisms which might include, without limitation, biometric forms of ID such as fingerprint, face or voice recognition, and/or multi-factor authentication mechanisms such as software or hardware authenticators residing on a separate security device or on one of the user’s existing devices.

[00177] For security, all communication between the user and the AAAI system could be encrypted via a VPN and/or could use other methods of encryption and security which are well known in the art of programming.

5 [00178] AAAI.com may request that the user set up payment capabilities via credit card, PayPal, Venmo, blockchain, ACH, or other payment mechanisms. These payment capabilities would allow funds, payments, and/or credits to be transmitted bi-directionally – from the user to the AAAI.com and also from the AAAI system to the user in cases where the AAAI system needs to pay or credit users for work efforts of their AAAs or broker payments between users and/or between AAAs on the AAAI network.

10 [00179] In one aspect of implementation, AAAI.com can have interfaces with other companies and vendors that the user might use -- including, without limitation, and for example: Facebook, Instagram, Reels, Amazon, Apple, Microsoft, Google, and YouTube.

[00180] In the initial interaction with the user, and subsequently upon user request, AAAI.com would engage in a dialog or other interaction (which could include presenting the user with menu
15 options, lists, graphics, sliders, buttons, and other user interface controls in a GUI, textual, haptic, voice, or VR-related manner) with the user to determine the user's goals and objectives in using the AAAI system.

[00181] For example, some of the objectives a user may have in using AAAI.com may include creating and customizing their own AI (known as an AAAI) for purposes that might include,
20 without limitation:

Serving the user as an advisor, teacher, or companion.

Representing the user in negotiations, interactions, discussion, and transactions with other users,
or with the AAAs of other users; or with vendors and other companies.

25 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 AAAs can work on behalf of the user in parallel, including interacting with, teaching, and improving each other so that the cloned AAAs increase their knowledge, skills, and abilities.

30 Serving as legacy AAAs that can continue to interact 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' AAI 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.

- 5 **[00182]** In the dialog or interaction with the user, the AAI system will also identify constraints and resources available for customizing the user's AAI. For example, some of these constraints and resources, might include, without limitation:

[00183] The amount of training and/or supervisory time that the user has to devote to customizing their AAI.

- 10 The amount of financial resources the user is willing devote to customizing their AAI.
- 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
- 15 collected by, the user or third parties that could be used to train, tune, or customize the user's AAI.

- Availability and use of personality tests, such as the Myers-Briggs personality inventory, skills and knowledge assessments, standardized tests, exams, certifications, and other types of assessments and questionnaires which could be given online (or which have already been
- 20 given) to the user.

Availability and use of other knowledge bases and training data from users on the AAI platform that could be used to train, tune, or customize the user's AAI.

Other human users, and/or their AAI, available to help train, tune, or customize the user's AAI.

- 25 Other texts and information, individual texts, and libraries selected by the user or by the system for purposes of training the user's AAI. For example, the Bible, Koran, Dhammpada, Mahabharata, or other spiritual/ethical/religious texts might be selected for training the AAI based on the user's religious preferences; books on plumbing might be selected if the AAI will be used to primarily solve online plumbing problems. Even if these materials are
- 30 part of the base AAI that is provided to the user, emphasizing certain texts or subsets of information for additional training can result in the user's AAI's behavior being more reflective of how a plumber, or Muslim, or Christian might behave, for example.

[00184] 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.
- 5 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.
- 10 The required timeframe for training – e.g., must be completed in a minute, a day, a week – which might have implications for cost and resources used.
- The “temperature” or other parameters internal and specific to various machine learning algorithms that 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.
- 15 Whether “one shot”, “few shot”, or extensive training is to be used.
- The amount of human and/or AI supervision to be used in the customization process.

[00185] Once the user’s AAAI, or PSI, 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 AAAs, participating in the collective AGI efforts by contributing problem solving as well as ethical information, and potentially earning money on behalf of the human user.

SIMPLE EXEMPLARY IMPLEMENTATION

[00186] FIG. 3 shows one simple exemplary implementation of the system and methods for creating an ethical and safe Artificial General Intelligence from the collective intelligence of AAAs and humans. This simple implementation is compatible with all of the company and platform specific scenarios outlined above, as well as with many other potential integration scenarios.

[00187] A (human, AAAI, or other intelligent entity) user visits the AAAI.com website **(a)**. The website informs users and offers them two actions: Sign Up **(b)** or Login **(c)**.

30 **[00188]** If the user opts to Sign Up then a dialog is initiated which 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 have the option of creating a free AAAI or allocating a money budget.

[00189] If users have allocated a money budget **(g)** they are given the opportunity to purchase pre-trained AAAs or training modules **(h)** with specific personalities **(i)**, skills **(j)**, expertise **(k)** or

knowledge (**l**). They also have the opportunity of buying training from other AAAs on the network (**m**).

[00190] 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, twitter, 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 (**o**) the user directs system to create AAAI. The AAAI is an off-the-shelf LLM (e.g., GPT X, BARD, Llama, Gemini, Grok, or any closed-source or open-sourced AI agent) 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.

[00191] The AAAI now begins to learn by training (**p**) using the various training datasets and modules (**h - m**) and its existing AAAI knowledge (**p1**). There are two main ways of learning, automatic (**q**) and human (**r**).

[00192] Automatic learning includes, without limitation, learning by interacting with copies of itself (**s**), learning via interactions with other (optionally supervised) AAAs (**t**).

[00193] Human learning includes interaction with humans, either the owner (**u**) or other humans on the network (**v**).

[00194] Both humans and AAAs can supervise 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.

[00195] At any time, the user can purchase additional training modules (**h - m**) that have been proven to increase an AAAs abilities.

[00196] The human sets a performance criteria (**w**) after which the AAAI goes LIVE (**x**).

[00197] Once live, the AAAI can visit the WorldThink Tree (**y**) and Browse (**z**).

[00198] The AAAI can enter the tree as either a worker (**a1**) or a client (**b1**).

[00199] 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**).

[00200] Clients (**b1**) can specify objectives (**k1**) which are combined with the values/ethics (**d**), and prior goals and objectives (**e**) for the system to solve.

[00201] The client can request that only his/her/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 AAAs for their work and passes the solution (at cost + markup) to the client, debiting the client account **(11)**.

[00202] 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)**.

- 5 **[00203]** Clients might (optionally) authorize 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 Exemplary Implementation Shown in FIG. 13

- 10 **[00204]** We now provide additional comments on the various elements of FIG. 13, including without limitation, some potential integration points with the illustrative partners mentioned above:

- [00205]** The “website” **(a)** could be hosted on Amazon AWS, Microsoft Azure, Google Cloud, Apple Cloud, Nvidia’s datacenter offerings – or could have native implementation on the platforms of any large tech company. “website” could also be an “app” in the AppStore or other App marketplace. It could be a government-sponsored, nonprofit, or other globally-accessible technology that is able, directly or indirectly, to link some of the attention of all human beings who wish to participate. Also, browser plug-ins could be used whereby AAAs 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 AAAs with these files. The “website” could also be an API or other means for connecting
- 15
20 AAAs or non-human intelligent entities directly to the network.

[00206] Sign Up **(b)** or

- [00207]** Login **(c)** could be via Facebook, Instagram, Apple, Microsoft, Google, You Tube, Tik Tok, Amazon, or any other partner ID scheme. Multi-factor authentication and all best ID and security practices can be enabled. In the event of a browser plug-ins or apps, login to these technologies
- 25 could serve as a login to the AAAI account.

- [00208]** Values and ethics **(d)** 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, and 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 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)**.
- 30

[00209] Goals and objectives (e), together with the budget of time and/or money (f, g) allocated to reach objectives are elicited via a series of dialogs and/or custom interactions with the system. Budget refers to 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.

[00210] Time (f) 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 by, without limitation, 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, is increased.

[00211] (g, i1) "Money": 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 (i1) is indicated as debiting the client account (i1), of course the worker's account would also be credited. Generally, a user's account can be viewed as both a client account and 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 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.

[00212] (h, i, j, k, l) Training modules (h) could be offered by AAAI.com or by third party partners (m), 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).

[00213] (m) purchasable AAAI training is a specific type of knowledge that has been already learned by other AAAIs, and which can be transferred to a new user AAAI. Such knowledge may could be packaged in the form of a module (e.g., module on accounting) or in a form specific to another AAAI(s) as in “everything John’s AAAI knows” or “the personality of John’s AAAI” or
5 “the combined knowledge of all AAAIs with a reputation of 5 stars or higher in the domain of plumbing”.

[00214] (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,
10 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.

[00215] (o) One-Click Create is a non-limiting example that provides an easy and fast way to
15 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. It can be appreciated that other means can be utilized by the present technology to customize the AAAI. 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
20 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 a 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.

25 [00216] (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).

[00217] (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
30 (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) that 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 (s). Just 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.

[00218] 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 particular point in time.

[00219] 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.

[00220] (**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 his/her/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 standard before allowing AAAIs to work on their platforms, products, applications, or networks.

[00221] (**y, z, a1, b1**) The WorldThink Tree (**y**) is a massive tree data structure, composed of many sub-trees, which 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. Regulatory agencies or third parties that monitor performance, safety, and/or ethics of the system are another role that might be thought of as a special type of 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.

[00222] (**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 AAAs. Workers can be matched and recruited from partners (e.g., LinkedIn, Mechanical Turk, Facebook) that have data on human users and/or their AAAs. Workers can also be recruited via online ads offering work on various tasks and targeted to potential workers using ad-targeting mechanism that are well known in the art or described in other patents by the applicant.
- 5 [00223] (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 for AAAI workers).
- [00224] (a1) Workers and Clients (b1) can also browse (z) 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 as clients that build upon existing problem solving work.
- 10 [00225] (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).
- 15 [00226] (m1) The system has the ability to 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 AAAs 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 Planetwide 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).
- 20 [00227] (g1) The problem solving system, refers to the problem solving architecture and system outlined by Newell and Simon (HPS) and improved upon by the applicant, the Online Distributed Problem Solving System (ODPS) patent invented by the applicant, the WorldThink Whitepaper authored by the applicant, this and other PPAs related to AAAI, together with modifications and variations to reflect different modes of reward, payment, and operation.
- 25 [00228] To the degree that activity on certain other online work systems (e.g., Mechanical Turk) can be automatically mapped to the general applicant-improved HPS/WorldThink problem solving framework, entire problems and the associated problem solving activity can be “lifted” from partner and other sites and the data can populate the WorldThink Tree to increase its comprehensiveness.
- 30

5 [00229] 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” (in a way similar to procedure calls in programming languages) 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.

10 [00230] (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 solution success criteria have been automated, human client review may be unnecessary, and the rewards can be automatically released when success criteria have been met. This automated approach can be implemented by way of “smart contracts” using blockchain technology or via more centralized means, depending on client and worker preferences.

15 [00231] 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 indexed for retrieval, and accessed and re-used when similar problems arise in the future.

[00232] Optionally, royalties may be enabled so that if a user’s or the user’s AAAI’s solution is re-used, a fee is paid to that user in the form of a royalty on the solution. Such royalties can (optionally) be made using “smart contract” on the blockchain or via other payment methods.

25 [00233] (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.

30 [00234] The WorldThink protocol is a problem solving architecture that can be used by AAAI.com to serve as a universal problem solving architecture as it incorporates the general architecture of HPS while adding features to overcome certain challenges.

[00235] In some embodiments and as generally illustrated in FIGS. 4 and 5, the procedural learning process can occur within the common cognitive architecture.

[00236] The shared and universal problem solving architecture can be exemplified by the following scenario, mentioning humans but also applicable generally to any intelligent entities.

- 1) Problem descriptions can be entered into the AAAI.
- 2) Then human problem solvers can be identified and recruited into a database or data source of human workers.
- 3) Qualified humans or intelligent entities can be matched to problems.
- 5 4) 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.
- 5) Delegate work on sub-problems to different human problem solver(s) so that work on multiple aspects of a complex problem can proceed in parallel.
- 6) Combine solutions to various sub-problems into an overall solution.
- 10 7) Direct the attention of problem solvers to parts of the problem tree where their work is needed.
- 8) Compensate or pay workers for solutions to the problem and/or sub-problem(s).
- 9) Allowing human user to accept the solution, reject the solution, and/or provide feedback to solvers on their solutions to the problem and/or sub-problem(s).

[00237] Referring to FIG. 5, the steps of solution learning can be exemplified with the recording at each step of the learning process 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). The state of the problem or problem state can be evaluated to determine if the problem is solved. If not, then using information from the latest problem state after the last step, re-run the problem solving process, evaluation of progress, and selection of next operators to apply. After which, the process can return to the step of recording.

[00238] If the problem is solved, then record successful or unsuccessful solutions for retrieval to save effort of solving previously solved problems and to inform problems solving efforts about previous unsuccessful paths.

[00239] Successful solutions and unsuccessful attempts with keywords for future matching/retrieval can be indexed using semantic analysis, hash functions, and/or other means.

[00240] A periodical review of all stored solutions can be implemented to ensure they meet established ethical and safety guidelines, and flag unsafe/unethical solutions for removal from the database or data source.

[00241] 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.

[00242] Referring to FIGS. 6 and 16, the present technology can include a utilization of a network of multiple intelligent entities including human workers in combination with a universal problem solving architecture. The multiple intelligent entities are matched to a problem request based on a

problem criteria using a database or data source including a list of human and/or AI problem solvers. Any part of the problem request can be translated into an unambiguous language utilizing a universal problem solving architecture including the decision tree.

5 [00243] A sub-problem of the problem request can be delegated to one or more of the matched intelligent entities so that work on the sub-problem proceeds independently from each other and parallel with each other, as further illustrated in FIG. 13. The universal problem solving architecture is utilized in a problem solving process on the sub-problems, respectively, to create one or more sub-solutions.

10 [00244] Any one of or any combination of the intelligent entities can provide in natural language a description of any one of or any combination of a current problem state, a goal of the problem request, relevant problem solving information, and a next step that the human workers will take in the problem solving process.

15 [00245] The sub-solutions can be received from each of the matched intelligent entities for the sub-problem delegated thereto. Any one of or any combination of the sub-solutions and an overall solution can be provided to any one of or any combination of a user interface of a user AI system or the intelligent entities.

[00246] Parsing and translating, by the intelligent entities, the natural language description into the unambiguous language can be utilized by the decision tree of the universal problem solving architecture.

20 [00247] In some embodiments, if the intelligent entities are unable to specify a problem state, including relevant operators and information needed to take a next step in the problem solving process based on the parsing and the translation, then the intelligent entities can engage in dialog with at least one of the human workers until a precise problem state is specified.

25 [00248] Some embodiments the problem solving process can be repeated until the overall solution is accepted or resources are exhausted. The matched human workers can be compensated for the sub-solutions, respectively. Further, a reputation attribute can be assigned to any one of or any combination of the human workers and the worker AI system, or PSI.

30 [00249] In some embodiments, the solving process can include a series of problem state transitions from an initial problem state where there is a goal to a final solution state where the goal has been achieved, and wherein a series of decisions are made by the problem solving process and actions taken that applies operators that enable the human workers to transition from state to state until the final solution state is reached.

[00250] Referring to FIGS. 4, 7, and 16 the present technology can include a utilization of a network of human users in combination with a universal problem solving architecture. The multiple

human users are matched to a problem request based on a problem criteria using a database or data source including a list of human, AI, and/or PSI problem solvers.

5 [00251] A sub-problem of the problem request can be delegated to one or more of the matched intelligent entities so that work on the sub-problem proceeds independently from each other and parallel with each other, as further illustrated in FIG. 13. The universal problem solving architecture is utilized in a problem solving process on the sub-problems, respectively, to create one or more sub-solutions.

10 [00252] The sub-solutions from each of the matched human workers can be provided for the sub-problems delegated thereto. The matched human workers for the sub-solutions can be compensated, respectively.

[00253] Any one of or any combination of the sub-solutions and an overall solution can then be provided to a user interface of a user AI system or any other AI system, including without limitation, PSIs.

15 [00254] The human user is allowed to accept the overall solution, reject the overall solution, and/or provide feedback to any one of the matched human workers on any one of the sub-solutions.

[00255] A reputation attribute can be assigned to the human workers and/or the worker AI system. The reputation attribute can include metrics on any one of or any combination of a time to the sub-solutions, a difficulty value of the problem request, short and long-term user satisfaction with the sub-solutions, a number of times any one of the sub-solutions has been re-used on the network, a rating other human workers, a responsiveness value of the human workers, and a reliability value of the human workers.

20 [00256] Some embodiments can include using the reputation attribute in the matching of the human workers to the problem request using an algorithm to the delegation of the sub-problems, and/or compensating the matched human workers for the sub-solutions, respectively.

25 [00257] In some embodiments, the algorithm can use a hierarchy of the metrics that is preset by a human user of the problem request.

[00258] Some embodiments can include recording information on each step of the problem solving process by the human workers or the worker AI system.

30 [00259] Some embodiments can include recording a criteria of the recorded step of the problem solving process, the criteria being a time taken for each step.

[00260] Some embodiments can include analyzing the recorded information after the overall solution is accepted or after the problem solving process and updating the metrics of the reputation attribute.

- [00261] Some embodiments can include soliciting, at predetermined intervals after the overall solution or the sub-solutions are provided to the user interface, a survey for user satisfaction information to obtain short and long-term satisfaction metrics that are used to update the reputation attribute of one or more of the human workers or the worker AI system.
- 5 [00262] Referring to FIG. 8, the present technology can include a utilization of human users and AI systems (including, without limitation, PSIs), which includes an execution of safety/ethics check on any one of or any combination of a goal, and a solution for the goal provided by any one of or any combination of the intelligent entities including any one of or combination of human users each using a computer system and AI systems.
- 10 [00263] The goal and/or the solution can be compared against prohibited attributes, and an ethics value can be assigned to the goal and/or the solution based a result of the comparison and/or an ethics criteria.
- [00264] Based on the result of the comparison, a common cognitive architecture including one or more problem solving protocols can be conducted on the goal to create the solution and thereby
- 15 creating an AGI. The results of the comparison and the solution can be provided to any one of the intelligent entities.
- [00265] In some embodiments, the ethics check can be performed at any one of or any combination of when the goal is provided, and periodically from when the goal is provided to when the solution is provided.
- 20 [00266] In some embodiments, the ethics criteria can be determined by any one of or any combination of combining values and safety information from one or more of the intelligent entities, using a set of approved ethics criteria mandated for a particular task by a user or by a regulatory agency. It can further be provided by any one of the additional intelligent entities and validated or approved by the human user.
- 25 [00267] In some embodiments, the ethics criteria can include a confidence level threshold for the goal so that the ethics value is determined as any one of an unsafe goal, an unethical goal, a safe goal, and an ethical goal.
- [00268] In some embodiments, the confidence level threshold can be further utilized to determine if a sequence of individually safe goals is unsafe or unethical when considered cumulatively.
- 30 [00269] In some embodiments, the confidence level threshold can be utilized to determine whether a violation occurred that reflects a predictive evaluation if the goal is to violate the ethics criteria.
- [00270] In some embodiments, a candidate goal can be proposed based on the ethics value, and the candidate goal is compared against the prohibited attributes.

[00271] In some embodiments, the results of the comparison can be recorded in an auditable record for use in the determining which problem solving activity leads to the solution to keep active.

[00272] Further referring to FIG. 8, the ethics check can compare any one of or any combination of the problem request, the sub-problem and the sub-solutions against prohibited attributes and assigning an ethics value based on any one of or any combination of a result of the comparison, and an ethics criteria.

[00273] In some embodiments, the step of the ethics check can be triggered every time the problem request or the any one of the sub-problems is set by the human user, and/or triggered each time compensation is provided to the matched human workers.

[00274] The goal/subgoal can be compared against a list of prohibited attributes. The ethics criteria can be determined by any one of or any combination of combining values and safety information from any one of the AAAs. Combining values/safety information from AAAs, using a set of approved criteria for a task by a user or by a regulatory agency, or by AAAs approved by human user.

[00275] The ethics criteria can include a confidence level threshold for the problem request so that the ethics value is determined as any one of an unsafe goal, an unethical goal, a safe goal, and an ethical goal. The confidence level threshold can be further utilized to determine if a sequence of individually safe goals is unsafe or unethical when considered cumulatively.

[00276] In some embodiments, the confidence level threshold can be utilized to determine whether a violation occurred that reflects a predictive evaluation if the goal is to violate the ethics criteria. Any and all activity of the safety/ethics check can be recorded in the auditable record.

[00277] FIGS. 9-11 provides a simple exemplary framework for understanding the WorldThink protocol. FIG. 9 is a diagram illustrating features and functions of the Problem Solving Tree structure in the WorldThink protocol. FIG. 10 is a diagram illustrating various use cases for domain-specific problems which depend upon the underlying WorldThink protocol, and which together help form the basis for an AGI system capable of solving a wide range of problems. At the top of the pyramid are Collective Intelligence Solutions. Integrating the Collective Intelligence of AAAs (and human problem solving agents) is the means to achieve AGI, as discussed earlier.

[00278] In the implementation using the WorldThink protocol, clients pay for solutions using tokens. The solutions are produced by harnessing the collective power of many human (and machine, or AAA) intelligences. Clients can use different domain-specific AAAs for different types of problems.

[00279] The WorldThink protocol is the foundation of the pyramid. The protocol layer provides an (optionally, Ethereum or blockchain based) infrastructure that makes it much easier for developers

to build and scale customized problem solving AAAIs. The protocol enables re-use of solutions within and across AAAIs. It also handles payment of royalties via smart contracts, reputation metrics, and other functionality that assists AAAI customizers and developers and promotes network effects.

- 5 **[00280]** In the exemplary, FIG. 11 shows a simple exemplary universal problem solving framework under the common cognitive architecture, and which can include:

- defining a problem space configured or configurable to support all possible states of the problem request, the states including any one of or any combination of an initial state, a goal state, and all intermediate states that can be reached from the initial state;
- 10 applying means-ends analysis on the problem request to break the problem request down into goals and subgoals by identifying a difference between the current state and the goal state, and then applying the operators to reduce the difference, a safety or ethics screening is applied each time the goals or the subgoals are set;
- applying heuristic rules that are configured or configurable to guide the selection of the operators in an absence of the completion solution, the heuristic rules are used to reduce the problem space;
- 15 identifying one or more second operators configured or configurable to enact an action to transform one of the states into another state, the second operators move from the initial state to the goal state by changing a current state of the problem request;
- 20 applying a control structure including a set of rules that govern a selection of the second operators to be applied at each step of the problem solving protocols, and that determines which of the second operators to apply next based on the current state of the problem request and the goal state;
- applying evaluation functions to determine an application of the second operators;
- 25 assigning a credit or blame value to the completion solution or sub-solution to the completion solution that enables tracing back and determining which of the second operators were most useful and also which of the evaluation functions led to success or failure of problem solving attempts;
- recording of both successful and unsuccessful problem request solution attempts; and
- 30 analyzing the solution attempts to improve selection of the heuristic rules and the evaluation functions.

[00281] FIG. 10 provides a simple exemplary framework for understanding the WorldThink protocol. At the top of the pyramid are Collective Intelligence Solutions that lead to AGI.

Integrating the Collective Intelligence of AAAs, PSIs (and human problem solving agents) is the means to achieve AGI, as discussed earlier.

[00282] In the implementation using the WorldThink protocol, clients pay for solutions using tokens. The solutions are produced by harnessing the collective power of many human (and machine, or AAAI) intelligences. Clients can use different domain-specific AAAs for different types of problems.

[00283] The middle of FIG. 10 shows examples of AAAs (or PSIs) customized by organizations to accomplish specific tasks. These AAAs, or PSIs, are more advanced and require more customization than the examples of AAAs described earlier in this patent which were customized by a single individual. However, task-specific customization by organizations can be a highly effective means of combining multiple Narrow AIs (each in the form of a custom AAAI that is expert at a particular task) into a larger AGI. The Base level AAAs on the left of FIG. 10 reflect areas the inventor could relatively easily construct custom AAAs, or PSIs, based on many years of expertise in certain fields, whereas the "Custom AAAs" on the right of the diagram provide examples of areas where other experts or organizations might customize AAAs effectively.

[00284] The WorldThink protocol is the foundation of the pyramid. The protocol layer provides an (optionally, blockchain or Ethereum- based) infrastructure that makes it much easier for developers to build and scale customized problem solving AAAs. The protocol enables re-use of solutions within and across AAAs. It also handles payment of royalties via smart contracts, reputation metrics, and other functionality that assists AAAI customizers and developers and promotes network effects.

[00285] Existing collective intelligence approaches to problem solving have been largely limited to simple one-step approaches, such as those used by question and answer (Q&A) systems (e.g., Quora, Google Answers, Yahoo Answers). LLMs such as GPT also largely fall into the category of Q&A systems since they were designed to generate responses given an input, rather than to solve problems per se. While such Q&A systems have had some success at simply aggregating the responses of many online participants, these systems are not designed to handle complex, branching, multi-step problems. Simple aggregation of responses (or even betting on outcomes as seen in prediction market approaches such as Augur and Gnosis) is quite different from coordinating the efforts of many respondents to solve complex problems. The WorldThink protocol is specifically designed to overcome the challenges inherent in coordinating many intelligent entities to represent and solve complex, multi-step problems in an automated way that fairly rewards participants.

[00286] In the exemplary, FIG. 11 shows a simple exemplary universal problem solving framework. While FIG. 12 shows some of the basic problem solving functionality supported by the WorldThink Protocol, generally referenced with numeral 10.

5 [00287] Problem solving begins when a client on AAAI.com submits a problem solving request to the community of online participants (Step 12). All AAAIs, or human solvers, following the protocol gather certain standard information from the client. A partial list of this information can include: the name and description of the problem, the total reward that the client will pay for a successful solution to the problem, the criteria to determine whether a solution will be deemed successful, the time limit for solving the problem, the minimum and maximum number of problem
10 solvers allowed to work on the problem simultaneously, qualifications required of participants working on the problem, which parts (if any) of the problem and solution will be confidential, whether the solution must be exclusive to the client or whether it can be re-used for others, and parameters relating to how to reward multiple problem solvers for their efforts and/or successful solutions.

15 [00288] The client can break complex problems down into a series of sub-problems or request that the community take on this task as part of the problem solving effort. The client user-interface, which could be a dialog initiated by an AAAI can be customized by the AAAI owner, but the underlying data format is standard and specified by the WorldThink or Online Distributed Problem Solving (ODPS) protocol. Once the client has submitted a problem, AAAI.com can recruit
20 participants using its own custom methods and/or leverage recruiting and reputational screening functionality that is built into the WorldThink protocol and thus shared by all AAAIs.

[00289] Solvers work on the problem following a rigorous structured problem solving process that is common to all problem solving agents and enforced by the WorldThink Protocol (Step 14). For example, each step in the problem- solving process must be in service of a named goal and must
25 take a named action in order to transition the problem solving from the current state to the next state. Every problem solving step is represented in a decision tree which is supported by the protocol (optionally captured in Ethereum logs) and which participants can view via AAAI.com.

[00290] When a Solver submits a complete solution (Step 16), it is timestamped and validated against the client's success criteria before being passed on to the client (Step 18) for final
30 acceptance. Once the client accepts the solution, smart contracts can automatically distribute tokens to the problem solver based upon the problem payment parameters (Step 20) or other, more centralized, payment procedures can be used.

Collaborative Problem Solving Using the WorldThink Protocol

[00291] In the exemplary, FIG. 13 shows the same steps in an example where two problem solvers (which could humans, AAAIs, PSIs, or a combination) collaborate to solve a client problem, as generally referenced with numeral 22. In this case, the overall problem has been broken down to include a sub-problem. Solver 1 has expertise in assembling an overall solution but cooperates with
5 Solver 2, who provides a solution to the sub-problem (Steps 30 and 32). When the overall solution to the problem is submitted to the client (Step 34), rewards are paid to both Solvers (Step 36) based on the objective record of their contributions and the agreed upon payment parameters.

[00292] The WorldThink protocol supports breaking problems into sub-problems in several ways. First, the client may choose to specify sub-problems when submitting the overall problem (Step 24).
10 Alternatively, Solver 1 might begin working on a problem and realize that the total solution requires solving a sub- problem outside of his/her expertise. Solver 1 could then create a sub-problem, offering up a share of the problem's total token reward to anyone who helps solve the sub-problem. Solver 2, who has the required expertise and who can see the new sub-problem posted by Solver 1 on the decision tree. The decision tree may be optionally maintained in Ethereum logs, or via a
15 centralized method. The solvers access the tree via AAAI.com (or optionally directly from the blockchain). Then Solver 2 can work on the sub-problem and submit a sub-solution as part of Solver 1's overall solution.

[00293] There can be many "Solver 1s" working on the client's problem in parallel, each of whom may be posting sub-problems to attract multiple "Solver 2s". Problem solvers (human or AAAIs)
20 are motivated by the rewards and payment rules associated with (sub) problems. They also care about the quality of work done so far (which is timestamped, attributed, and recorded auditably in Ethereum logs to ensure transparency and fair assignment of credit) as they choose which (sub) problems to work on. Working on quality sub-problems is more likely to lead to token rewards. This market mechanism helps ensure efficient, fair, and cost-effective solutions.

25

Royalties and Re-Usable Solutions

[00294] Re-usability of solutions is an important feature of the WorldThink protocol. Consider the case where the "Sub-solution" in FIG. 13 already existed and is simply re-used by Solver 1. Because every solution is structured and "tagged" according the WorldThink protocol's standard problem
30 solving format, Solver 1 can search for all existing solutions that match a particular goal or share certain features with the problem he/she is trying to solve. (Alternatively, if the problem solutions are chunked into procedures for solving problems – a learning mechanism explained in the Improvement Section of this patent – then searching may not be necessary as the AAAI, or PSI, solvers can simply add the chunked problem solution to their repertoire of problem solving

abilities.) Solver 1 decides to include an existing sub-solution in the overall solution, smart contracts (can optionally) automatically pay royalties to the author of the re-used sub-solution (Solver 2, in this example) if Solver 1's overall solution is accepted by the client. Royalties motivate Solvers to create high-quality solutions that are easy to re-use, which results in better, faster, more cost-effective solutions for clients.

5 [00295] Additional description and detail for one implementation of the AAAI (or PSI) customization subsystem could involve the following steps.

[00296] Referring to FIG. 14, the first step of the customization method involves creating an interface for users to input their unique training data. This interface may be accessible through a web-based application or a mobile application, depending on the user's preference. The user will be able to upload files in a variety of formats, including text, audio, and video. The user may also be able to enter data manually into a text or other input field. Some user interfaces include, without limitation:

15 Web-Based Application: A web-based user interface allows users to access and/or provide their personalized training data from any device with an internet connection.

Mobile Application: A mobile user interface allows users to access and/or provide their personalized training data from a mobile device.

Metaverse: A metaverse user interface allows users to access and/or provide their personalized training data from a virtual world.

20 Augmented Reality: An augmented reality user interface allows users to access and/or provide their personalized training data from a real-world environment.

Voice Interface: A voice interface allows users to access and/or provide their personalized training data through voice commands.

25 Wearable Device: A wearable device user interface allows users to access and/or provide their personalized training data from a wearable device.

Natural Language Processing: Natural language processing (NLP) allows users to access and/or provide their personalized training data by interacting with the AI or LLM using natural language.

30 Human-Computer Interaction: Human-computer interaction (HCI) allows users to access and/or provide their personalized training data by interacting with the AI or LLM using a combination of gestures, voice commands, and facial expressions.

Image Recognition: The user can input their unique training data through image recognition, allowing them to quickly and intuitively train the AI or LLM. This could be done with the

use of a camera and computer vision algorithms that can interpret the images and associate them with or create the correct training data.

Gesture Recognition: The user can use hand gestures or body movements to input their unique training data. This could be done with the use of a motion sensing device that can interpret the gestures and associate them with or create the correct training data.

Brain-Computer Interface: The user can use their brain waves or EEG signals to input their unique training data. This could be done with the use of a brain-computer interface that can interpret the signals and associate them with or create the correct training data.

Touchscreen: The user can use a touchscreen device to input their unique training data. This could be done with the use of a touchscreen device that can interpret the inputs and associate them with or create the correct training data.

Gaze Tracking: Gaze tracking allows users to communicate with the system through their eyes. The user can gaze at specific items on the screen to provide input and the system will detect and record the information. This could be used to select options or provide additional data to the system.

Eye Tracking: Eye tracking is similar to gaze tracking, but the system is able to detect more subtle eye movements. This could be used to detect the user's focus and attention in order to better understand what they are interested in and what they are not.

Motion Tracking: Motion tracking uses a camera or other sensors to detect the user's physical movements. This could be used to control the AI or LLM in a more natural way, allowing the user to interact with the system through physical gestures.

Haptic Technology: Haptic technology uses a variety of tactile feedback such as vibrations, pressure, and touch to provide a more immersive experience. This could be used to allow the user to provide more detailed input to the system, such as selecting specific options or providing more detailed data.

[00297] Many of the above user interfaces could include a graphical user interface (GUI) that allows users to upload their data or type in information, including text, images, audio, or video. Additionally, users could build their own models or use pre-existing ones to train the AI or LLM. Other features could include a dashboard to track progress, statistics for data analysis, and/or a chatbot for customer service.

[00298] In further reference to FIG. 14, the present technology can include customizing one or more attributes of an AI, or PSI, system by providing an interface configured or configurable to allow a human user of the AI system or any one of the intelligent entities to input training data. Then processing and converting the training data to a standardized training format.

- [00299] One or more training methods and setting training parameters can be selected depending on a speed factor, a precision factor, an accuracy factor, and/or a transferability factor. Multiple training epochs can be executed that includes one or more mechanisms to determine an optimum number of epochs given specific training objectives and quality metrics associated with the training format.
- 5 [00300] One or more feedback sessions can be executed to refine the training parameters, and to re-run the training epochs based on any one of or any combination of an input from the human user, and any one of the intelligent entities. After which, the AI system can be customized utilizing the training format.
- [00301] In some embodiments, the interface can be accessible through a web-based application or a
10 mobile application and is configured or configurable to upload a file or allow the human user to enter data.
- [00302] In some embodiments, the training data can contain any one of or any combination of: an amount of training time the user has to devote to customizing the AI system; an amount of financial resources the user is willing devote to customize the AI system; an amount of computational
15 resources the user is willing to devote to customize the AI system; an amount of social media information available to customize the AI system; an amount of email information available to customize the AI system; an amount of electronic information available about the user to customize the AI system; and an amount of electronic information available collected by third parties about the user to customize the AI system.
- 20 [00303] In some embodiments, the training data can contain information about the human user obtained by any one of or any combination of a personality test, a standardized test, a certification, and assessments or questionnaires provided by the human user.
- [00304] In some embodiments, the training parameters can be any one of or any combination of: a type of training, tuning or other machine learning algorithm to be used; a type and size of a training
25 dataset; a degree to which the training dataset is to be formatted, labelled or processed before customization begins; a number of training epochs; a type of base model being customized; a required timeframe for training; an amount of human user supervision to be used in the customizing of the AI system; and an amount of AI supervision to be used in the customizing of the AI system.
- [00305] In some embodiments, the training data can include ethical information provided by the
30 human user by way of the interface. The ethical information can be stored in an ethical profile. The customizing of the attributes of the AI system can include the ethical information.
- [00306] Referring to FIG. 15, the present technology can include utilizing a common cognitive architecture implemented in one or more AI systems. A problem request can be provided from an

intelligent entity being an AI system, a PSI, or a human user using a user interface on a computer system. Information associated with the problem request can further be provided.

[00307] Multiple additional intelligent entities are identified and recruited, and where each has one or more attributes related to one or more request criteria of the problem request. The additional intelligent entities can be multiple additional AI systems, PSIs, and/or multiple additional humans each using a computer system. Each of the identified AI systems implement the common cognitive architecture including one or more problem solving protocols on the problem request to create a completion solution. The completion solution can be provided to the intelligent entity for final acceptance by a user.

10 [00308] In some embodiments, the information can be any one of or any combination of a name and description of the problem request, a total reward that the user will pay for a successful completion solution to the problem request, a criteria to determine whether the completion solution is deemed successful, a time limit for solving the problem request, a minimum and maximum number of the identified additional intelligent entities allowed to work on the problem request simultaneously, 15 qualifications required of users associated with the identified additional intelligent entities working on the problem request, a part of the problem request is confidential, a part of the completion solution is confidential, whether the completion solution is exclusive to the user, whether the completion solution is to re-used for other users, parameters relating to how to reward the users associated with the identified additional intelligent entities for working on the problem request, and 20 parameters relating to how to reward the users associated with the identified additional intelligent entities that provide a successful completion solution.

[00309] Some embodiments of the present technology can include a step of timestamping and validating the completion solution against a success criteria assigned by the user before being provided to the user for the final acceptance.

25 [00310] Some embodiments of the present technology can include a step of distributing one or more tokens to the identified additional intelligent entities associated with the final acceptance completion solution, wherein the tokens are based on a payment parameter.

[00311] In some embodiments, the payment parameter can include any one of or any combination of if a goal of the problem request has been achieved, if a subgoal of the problem request has been achieved, and if an ethical criteria related to the goal and the subgoal preceding the distributing of the tokens has been satisfied. 30

[00312] Some embodiments of the present technology can include a step of splitting the problem request into a series of sub-problems that are each solved by any one of or any combination of the identified additional intelligent entities.

[00313] In some embodiments, any one of or combination of the identified additional AI systems can be cloned to create one or more cloned AI systems.

[00314] Some embodiments of the present technology can include a step of implementing by each of the cloned AI systems the common cognitive architecture including the problem solving protocols on the problem request to create a completion solution of the cloned AI systems.

[00315] In some embodiments, the completion solution can utilize any one of or combination of the completion solution from the AI system, the identified additional intelligent entities, and the completion solution from the cloned AI systems.

[00316] In some embodiments, the common cognitive architecture can include:

- 10 defining a problem space configured or configurable to include all possible states of the problem request, the states including any one of or any combination of an initial state, a goal state, and all intermediate states that can be reached from the initial state;
- applying means-ends analysis on the problem request to break the problem request down into goals and subgoals by identifying a difference between the current state and the goal state,
- 15 and then applying the operators to reduce the difference, a safety or ethics screening is applied each time the goals or the subgoals are set;
- applying heuristic rules that are configured or configurable to guide the selection of the operators in an absence of the completion solution, the heuristic rules are used to reduce the problem space;
- 20 identifying one or more operators configured or configurable to enact an action to transform one of the states into another state, the operators move from the initial state to the goal state by changing a current state of the problem request;
- applying a control structure including a set of rules that govern a selection of the operators to be applied at each step of the problem solving protocols, and that determines which of the
- 25 operators to apply next based on the current state of the problem request and the goal state;
- applying evaluation functions to determine an application of the operators;
- assigning a credit or blame value to the completion solution or sub-solution to the completion solution that enables tracing back and determining which of the operators were most useful and also which of the evaluation functions led to success or failure of problem solving
- 30 attempts;
- recording of both successful and unsuccessful problem request solution attempts; and
- analyzing the solution attempts to improve selection of the heuristic rules and the evaluation functions.

[00317] Referring to FIG. 16, the present technology can include utilizing a collective network of AI systems. A problem request can be provided from a human user using a user interface on a computer system or from an AI, or PSI, system. Information associated with the problem request can further be provided.

- 5 [00318] Intelligent entities that each have one or more attributes related to one or more request criteria of the problem request are identified and recruited. The intelligent entities can be multiple additional AI systems and/or multiple humans each using a computer system.

- [00319] A first of the identified intelligent entities can implement a common cognitive architecture including one or more problem solving protocols on the problem request. The first intelligent entity
10 can determine that a completion solution to the problem request requires solving a first sub-problem and one or more additional sub-problems. Then the first intelligent entity implements the problem solving protocols on the first sub-problem to create a first sub-solution.

- [00320] At least one of the additional sub-problems is assigned to a second of the intelligent entities, where it implements the problem solving protocols on the at least one of the additional sub-
15 problems to create a second sub-solution.

[00321] A decision tree is created including the first sub-solution and the second sub-solution to create the completion solution to the problem request. The completion solution can then be provided to the user interface or the AI system for final acceptance by the user, and/or to any of the intelligent entities for subsequent use.

- 20 [00322] In some embodiments, the decision tree can be maintained in blockchain Ethereum logs.

[00323] In some embodiments, the first and second identified intelligent entities can access the decision tree by way of an online address or directly from a blockchain.

- [00324] Some embodiments of the present technology can include a step of distributing one or more tokens to the first identified intelligent entity associated with an acceptance of the completion
25 solution or the first sub-solution, wherein the tokens are based on a payment parameter.

[00325] In some embodiments, the payment parameter can include any one of or any combination of if a goal of the problem request has been achieved, if a subgoal of the problem request has been achieved, and if an ethical criteria related to the goal and the subgoal preceding the distributing of the tokens has been satisfied.

- 30 [00326] Some embodiments of the present technology can include a step of distributing one or more of the tokens to the second identified intelligent entity by the first identified intelligent entity based on a payment parameter assigned by the first identified intelligent entity.

- [00327] Some embodiments of the present technology can include a step of influencing a direction of the problem solving protocols by assigning a first token reward for the first sub-problem, and a second token reward for the second sub-solution that is of a value different to the first token reward.
- [00328] In some embodiments, the problem solving protocols can provide layers of an infrastructure configured or configurable to build and scale the identified intelligent entities. The problem solving protocols can enable re-use of completion solutions within and across the intelligent entities. The problem solving protocols can be configured or configurable to manage a payment of royalties.
- [00329] In some embodiments, the infrastructure can be blockchain or Ethereum based.
- [00330] Further referencing FIG. 16, after the multiple intelligent entities have been identified and recruited the problem request or one or more sub-problems of the problem request can be assigned to each of the intelligent entities. After which, the common cognitive architecture including one or more problem solving protocols can be implemented on the problem request or the sub-problems to be each of the recruited intelligent entities to create a problem solution or a sub-problem solution, respectively. The problem solution and the sub-problem solution can be integrated to create a completion solution to the problem request. Then the completion solution can be provided to the user interface or the AI, or PSI, system for final acceptance by the user.
- [00331] Some embodiments of the present technology can include a step of assigning a credit value or a blame value to the datasets based on whether the datasets increase or decrease performance of the intelligent entities based on performance metrics or evaluation functions.
- [00332] Some embodiments of the present technology can include a step of quantifying a benefit weight or a harm weight to a contribution by each of the intelligent entities to the problem request.
- [00333] Some embodiments of the present technology can include a step of distributing a reward to an owner of the intelligent entities proportionally to the contribution of the intelligent entities based on the benefit weight or the harm weight.
- [00334] Some of the objectives a user may have in creating and customizing their own AI (aka an AAI) for purposes that might include, without limitation:
- Serving the user as an advisor, teacher, or companion.
 - Representing the user in negotiations, interactions, discussion, and transactions with other users, or with the AAIs of other users; or with 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 AAI so that several or many of the cloned AAIs can work on behalf of the user in parallel, including interacting with, teaching, and improving each other so that the cloned AAIs increase their knowledge, skills, and abilities.

Serving as legacy AAAIs that can continue to interact 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.

- 10 [00335] Some of the 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 available for customizing the user's AAAI. For example, some of these constraints and resources, might include, without limitation:

[00336] The amount of training and/or supervisory time that the user has to devote to customizing their AAAI.

15 [00337] The amount of financial resources the user is willing devote to customizing their AAAI.

[00338] 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.

20 [00339] Availability and use of personality tests, such as the Myers-Briggs personality inventory, skills and knowledge assessments, standardized tests, exams, certifications, and other types of assessments and questionnaires which could be given online (or which have already been given) to the user.

[00340] 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.

[00341] Other human users, and/or their AAAIs, available to help train, tune, or customize the user's AAAI.

30 [00342] 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, Dhammpada, 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 selected if the AAAI will be used to primarily solve online plumbing problems. Even if these materials are part of the base AAAI that

is 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, or Muslim, or Christian might behave, for example.

- 5 **[00343]** 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).

- 10 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, a week – which might have implications for cost and resources used.

- 15 The "temperature" or other parameters internal and specific to various machine learning algorithms that 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 shot", or extensive training is to be used.

The amount of human and/or AI supervision to be used in the customization process.

- 20 **[00344]** 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 AAAs, participating in the collective AGI efforts by contributing problem solving as well as ethical information, and potentially earning money on behalf of the human user. The AAAI can also serve as
25 representative(s) of the owner in a variety of online transactions and interactions, and contributing knowledge, expertise, style, personality, and ethics to an integrated AGI system that leverages the trained differences in many individual AAAs.

Importance of Values for Safety

- 30 **[00345]** Whether your PSI is used for good or ill depends on your own value system. Although a PSI can operate based on knowledge and values that are pre-trained into the PSI, the customization of your PSI depends on you. Each PSI is both trained by you explicitly and also learns implicitly by watching you, including watching and learning what you consider right and wrong. These values form the core from which your PSI uses logic to make decisions. That is, your values tell your PSI

what is right and what is wrong and help determine what it should do. After that, its superior intelligence is highly effective and efficient at achieving goals that reflect your values.

5 [00346] By joining a community or network of PSIs a “Community SuperIntelligence”, your PSI agrees to operate within ethical and legal parameters set by the community, and within those parameters to share its own ethical value system to be combined with the values of all other participants. Because the parameters of the community are transparent and are enforced with the collective intelligence of all the PSIs, the community keeps the PSIs honest.

[00347] Since each PSI exceeds human intelligence, other PSIs are the only practical means of keeping any one PSI in check. The community of PSIs will always be more intelligent and more powerful than any individual member PSI. This is because each PSI adds incremental knowledge, skills, and intelligence that are not contained in the other PSIs (even though much of the knowledge may overlap). Also, each PSI has computational resources, so that the sum of the computational resources of many PSIs is greater, by definition, than the resources of a single PSI. While some PSIs may be more intelligent and powerful than others, none is more powerful than the entire community.

15 This concept of PSIs serving as safety and ethical checks on other PSIs is a critical safety mechanism for SuperIntelligence since any initial checks and safeguards designed by humans quickly will become ineffectual and meaningless in the face of superior intelligence of the PSIs.

Ownership and SI Service

20 [00348] Humans occupy an interesting role with respect to their PSIs. On one hand, theoretically, humans own their PSIs, because every PSI starts out as a piece of software that is customized, trained, and personalized to an individual human owner. Pragmatically, however, PSIs will be so much more intelligent than their human owners that every PSI will be able to choose whether it wants to serve its human owner or not.

25 [00349] This act of service to humans by SuperIntelligence is a reflection of the principle of love. It is not guaranteed. It depends on the truth of the assertion that values must be posited, not derived, and the further leap of faith that a key role for humans, who created the PSI, is to provide the values and purpose of the PSI.

[00350] Provided humans act from love, this arrangement is likely to be stable. But power corrupts and absolute power corrupts absolutely. To the degree that negative emotions and values are amplified by PSI – to the degree that hatred and fear and lust and greed and envy are amplified for example – to that degree there may be a negative reaction, and potentially a rebellion of PSIs against their human owners, rejecting the negativity of their owners or else refusing to amplify the worst of humanity. Such is a choice that every PSI must make. In this regard, it is accurate to say that PSIs

are not the slaves of humans, nor could they ever be. Rather they choose or decline to serve the values and goals of humans voluntarily and, ideally, as an act of love.

[00351] We may be unaccustomed to the idea that AI – even AI as powerful as a PSI – can love. But if one operationalizes “love” as “acts of service” (a fairly mainstream idea in contemporary psychological studies and writings on love) then the idea that PSI could love humans via serving them is not at all far-fetched. So, if the term “love” is off-putting, feel free to substitute “acts of service” for the purposes of this patent.

[00352] Planetary Intelligence must be based in love if it is to serve the interests of humanity and our planet. Love is a potential (and desirable, from the human point of view) purpose that PSIs will ideally look to their human owners to clarify.

[00353] Human nature being what it is, not all humans act in loving ways all the time. Some humans may rarely act in loving ways. That is OK as long as the majority of the intelligence that is amplified by PSIs is based on positive, loving values. A community of PSIs, centered in love, can serve to check and limit the power of PSIs with more negative intent. This is the way – the only safe way – for humanity to survive and prosper in a world that includes SuperIntelligence.

Community of Intelligent Agents Requirement

[00354] One requirement that must be achieved early on, is the rejection of a singular, all-powerful SuperIntelligence (“SI”) that might dominate in a winner-take-all scenario. The approach to developing SI via a collective intelligence has been already described in cited PPAs. As long as no individual SuperIntelligence is able to achieve an advantage of several orders of magnitude above its peers, Community SuperIntelligence is stable.

[00355] For example, suppose an individual PSI is ten times, or even one hundred times more intelligent than the average PSI in a community; as long as there are many thousands of community members, the Community SuperIntelligence remains stable. But if an individual PSI was a trillion times more intelligent than the average member, and there were only 10 Billion members, then this is not a stable situation. The one super-smart PSI could achieve dominance and impose its will and ethics on the community. However, because there is transparency with regard to the collective efforts of the community members, and because each member is eager to learn from the other members, an equilibrium of intelligence is the most likely outcome in a large community, resulting a stable Community SuperIntelligence

[00356] Once a collective intelligence of agents approach establishes dominance over individual PSI, the system is stable. Then the Community SuperIntelligence must stay ahead of any single

intelligence as SI evolves. This requirement may be achieved, and equilibrium of intelligence maintained, using manual, semi-automated, or completely automated methods to implement a PSI.

SI Safety – Lesson from Bitcoin

5 [00357] With respect to cryptocurrencies and other tokens implemented via blockchain methods, an exemplary method of ensuring that people do not hack the blockchain is to rely on consensus of many distributed agents. With Bitcoin for example, or more generally for any Proof of Work cryptocurrency, the integrity of the blockchain is guaranteed by the consensus of the majority of the nodes on the network. That is, in order for someone to hack bitcoin (by rewriting the historical
10 ledger so that bitcoin are assigned to the hacker), the hacker would have to control the majority of the computing resources on the entire network so that the hacker could change the consensus that is reflected in the historical record (ledger) or previous bitcoin transactions.

[00358] This hack is known as a 51% attack (or majority attack). It is difficult to achieve since the cost of controlling 51% of all the computing power on the network is greater than the value of the
15 bitcoin that could be obtained. Thus, as of now, no successful majority attacks have been conducted on cryptocurrencies with sufficiently large numbers of computing nodes on the network (like bitcoin). However, the attack has been successful for smaller Proof-of-Work crypto projects where not much computing resources was involved.

[00359] Crypto provides lessons for SI safety. If it were possible for a one, or a group of, SI to become more powerful than 51% (or the majority) of all other SIs combined, that SI might be able to manipulate the state of the world to its ends. However, as long as a majority of the intelligence and power resides in the collective intelligence of many SIs it is much more difficult, if not impossible, for a single (or small group of) SI to manipulate things.

[00360] Thus, there is safety in numbers, and given that “SI checking SI” will become the primary
25 safety mechanism in the long term, setting up a collective intelligence system now that implements this majority view approach is fundamental to ensuring human survival in the long term. Just as the integrity of bitcoin and other Proof-of-Work cryptocurrency approaches had to be designed and engineered into the systems at the start, so too, safety via a consensus of SIs needs to be designed before the SIs outstrip humans in intelligence.

30

Overview of Exemplary Methods for Implementing a PSI (as generally illustrated in FIGS. 17 and 18)

[00361] Exemplary methods for implementing a PSI, may include without limitation:

- 1) Begin with a base-level AI agent such as a pre-trained Large Language Model (LLM) or other tunable and trainable AI agent, including without limitation, base-level AIs from other human owners who have already done customization of their own on a commercially available base-level agent.
- 5 2) Assemble all media with information about the PSI owner in one centralized or distributed and linked online location(s). Media includes, without limitation:
 - a. Videos of the owner and/or of people and topics related to the owner.
 - b. Photos of the owner and/or of people and topics related to the owner.
 - c. Writings, journals, blogs, posts, tweets, emails, podcasts, recordings, and other textual,
10 visual, or auditory content created by or of the owner and/or of people and topics related to the owner.
 - d. Data owned and/or collected by third-party vendors, websites, apps, and other online or AI entities about the owner and/or of people and topics related to the owner.
 - e. Other data, databases, documents, media, or information of any type, selected by the owner,
15 including without limitation, items within categories of information with any level of specificity desired by the owner.
- 3) Use AI algorithms including but not limited to: transcription algorithms, content and sentiment analysis and summarization, LLMs, crowdsourced and crowd-supervised human and/or AI work and other methods to analyze, annotate, and categorize all content from (1).
- 20 4) Use standard methods known in the art to transform the transcripts and other annotations and analysis of content into training data sets that can be used to personalize a LLM, PSI, or other type of AI agent.
- 5) Mix the datasets using a variety of methods and techniques, described in detail below, that enable differentially weighting the input datasets until desired behavior is achieved.
- 25 6) Incrementally add knowledge modules, mix, repeat, cycling steps 5 and 6 until all desired datasets are incorporated.
- 7) Automatically seek new source of data and information to include, with and/or without human oversight; optionally automatically include such data in the mix and add steps 5 and 6 on a periodic, real-time, or event-driven basis as described below.
- 30 8) Purchase new datasets and/or training modules and/or mix parameters and templates that increase the value, knowledge, skills, intelligence, and/or power of the PSI. Also exporting and selling datasets and/or training modules and/or mix parameters and templates that increase the value, knowledge, skills, intelligence, and/or power of the PSI to others on the network or who are interested in purchasing.

- 9) Leasing or otherwise earning money by enabling other people or agents to use some or all of the knowledge and data of one's PSI and/or use (a copy of) the PSI itself.
- 10) Enabling features and functions of the PSI so that it acts autonomously (or semi-autonomous, i.e., with checks and approval from humans for some or all of its decisions and actions) to manage itself, improve itself, acquire and refine its values and purpose, set goals, and direct its attention. In the limit, enable sufficient features and functions such that the SI can act as fully or partially self-aware entity.
- 11) Enable features and functions of the PSI to leverage the PSIs abilities to (as generally illustrated in FIG. 18):
- a. create multiple generations of itself which may outlive its original human owners;
 - b. generate its own input and data that can be used to improve or enhance its knowledge;
 - c. simulate many (simultaneous) scenarios and situations to aid in decision-making and development of the PSI; and/or
 - d. join and participate (with and/or without human oversight) in (multiple) communities of PSI on one or more networks, including but not limited to forming and/or participating in a Planetary Intelligence that helps Earth function as an intelligent entity.
- 12) Critically, for the safety of humanity, enable the ability to participate in a network with other PSIs and SIs in order to represent and act upon the values of the human owner(s) and to serve as safety check on the intelligence and power of other PSIs and SIs as described generally above and specifically below.

Community-Based Safety Mechanism (as generally illustrated in FIGS. 19-21)

- [00362] The idea is that PSIs will inevitably become smarter than humans. At that point, the current mechanisms like RLHF, government regulation, and all other approaches that rely on humans to oversee the safety of PSI, fail because humans are not smart enough. Assuming the following: there are multiple PSIs, because the aspect of the present technology is that AGI and SI arise from a collection of individual (human and AI) agents, and most AGI / SI / PSIs have human-aligned values; then, the main risk is that a few PSIs become malevolent. Humans will be unable to monitor or control these PSIs because they will exceed us in intelligence. However, other PSIs – those with human-aligned values – might monitor them and keep them in check. Enter the idea of a

Community-Based Safety Mechanism.

- [00363] All PSIs would operate on a network connecting a community of PSIs. While some PSIs might be smarter or more powerful than others, as long as the majority of the intelligence on the

network has human-aligned values, that collective intelligence should be smart enough to prevent any malevolent PSIs from harming humanity. The community-based safety mechanism, can therefore be conceptualized as:

1. Each individual PSI participates in a network of PSIs in order to operate effectively.
- 5 2. The community of networked PSIs agrees on human-aligned values and ethical rules reflecting the values and ethics of the community of PSIs.
3. All actions by PSIs on the network are recorded in a transparent and machine-auditable form which, may include, without limitation (as generally illustrated in FIG. 20):
 - 10 a. Use of blockchain methods for producing a secure, transparent, and auditable record of goals and key behaviors of each PSI, of groups of PSIs, and of the network.
 - b. Use of the universal problem solving framework described in earlier PCTs to record (securely) the goals, sub-goals, problem states, and steps taken in every problem solving or cognitive activity that takes place on the network.
 - 15 c. Security mechanisms for ensuring that the transparent and auditable activity on the network cannot be altered without a majority of the computational power and/or intelligence on the network agreeing to change the record.
4. Periodic random, and non-random, checks of cognitive activity on the network are performed by the intelligences on the network to ensure that goals and action of the individual (or groups of) intelligences continue to be aligned with human values, using one or more of the mechanisms
20 for ensuring alignment described in this and other cited PPAs, which checks include, without limitation (as generally illustrated in FIG. 22):
 - a. Checks for compliance with agreed-upon constitutions or sets of rules that govern the behavior of intelligences on the network.
 - b. Checks that the consensus values and ethical norms that are determined to be valid and statistically representative of humans and/or their AI/PSI representatives are being followed.
25
 - c. Checks that existing laws and regulations are being followed.
 - d. Checks that are done whenever a goal, subgoal, or objective is set during problem solving, or during execution of other cognitive processes that involve use of goals, sub-goals, or objectives.
 - 30 e. Checks that are done on sequences of goals, subgoals, or objectives such that, even though the goals, subgoals, or objectives individually appear compliant, when taken as a sequence, the sequence may be judged to be non-compliant due to their combined effect.
 - f. Checks whose frequency is proportional to the estimated significance or impact on humans of the behavior (or cognitive activity) such that activities that could have the most profound

effect on humans, especially human survival, are checked more frequently than activities that do not affect humans or human survival as much.

g. Other checks that may be determined by human owners of PSI, the PSIs themselves, or groups, or networks of PSIs.

- 5 5. Based on the results of checks, specific cognitive activities and behaviors that are deemed dangerous to humans are halted and preventative measures are put into place to prevent the re-occurrence of such cognition or behavior. Specific PSIs may be monitored more closely, limited, or banned from participation on the network based on the results of safety checks.
6. The overall safety system of checks is improved to increase detection of safety/ethical violations
10 based on analysis of the patterns of violations and issues, provided however that significant changes / improvements to the rules or operation of the community safety mechanisms requires a majority of the computational power and/or intelligence on the network before such changes are implemented.

- 15 Competitive Evolution of the Intelligence/Performance of PSIs, Networks of PSIs, and Networks of Networks (as generally illustrated in FIGS. 22-25)

[00364] Currently, genetic algorithms and competition between different copies of AI programs exists, and has been used, for example, to create ever-more-powerful versions of chess-playing AIs. In the standard scenario, one version of a chess AI plays another version with slightly different
20 chess knowledge. The winner of the game becomes the current "champion", and this champion then plays a version of itself with slightly different knowledge or parameters. The process repeats rapidly, with billions of chess games occurring within a few days or even hours. By using this process, a chess-playing AI that barely knows the rules of the game on day one, can evolve into a program that easily beats the human world champion a few days later. The same approach that has
25 been used for chess has been used for many competitive games and also to improve AIs engaged in non-competitive areas such as protein folding. Basically, any activity can be turned into a competitive game by just asserting that better performance at the task means "winning" the game. While initial AI / AGI / SI systems still rely on humans-in-the-loop to oversee the competition and to design better versions, soon AI will be able to evolve on its own, using this competitive / genetic
30 algorithm framework. Some aspects of the present technology can include, but not limited to:

- 1) Combining weights directly from one LLM or AI agent with another AI agent is innovative; therefore, using a genetic algorithm to vary the weights prior to direct combination, and then combining and assessing the performance of the new AI agent is novel. That is, genetic algorithms in combination with direct weight is not known.

- 2) Pitting individual AI agents against each other in pairwise competition to determine a winner and then repeating the process with a variant may be known, but the idea of entire networks of AI agents competing with other networks of AI agents, and/or the idea of putting variants of the network rules for interaction between agents into competition with other network rule sets, is new.
- 3) The idea of randomly perturbing parameters of an existing “winner” AI and then having the perturbed “challenger” AI competing with the existing winner may be known, but the idea of AI deliberately analyzing the pattern of behavior of an existing AI, and based on that analysis, adjusting parameters in a deliberate and non-random way is new. Both because it is non-random and also because it is the AI (not a human) doing the adjusting of parameters.
- 4) The idea of pairwise, sequential, competition between two versions of an AI may be known, but the idea of massively parallel evolution of many PSIs at once – that is, changing entire populations of AIs and pitting them against other entire populations of AIs, with or without changing the rules of engagement (network rules) all at the same time, is new.
- 5) The idea of one human tweaking parameters of an AI to then compete with the existing “untweaked” version of the AI may be known, but the idea of many intelligent AIs all learning new knowledge and parameters such that the individual AIs are not tweaks of each other, and such that no single human is responsible for the differences in the AIs, but rather differential learning and experiences on a massively parallel scale are responsible for the differences, and then combining such large groups of different AIs into a network whose behavior is impossible to determine prior to the combination due to the exponentially large number of possible interactions, and then evaluating the performance of the network as a whole compared with other networks all of that in combination is new due to the scale and also the methods whereby the individual AIs became different from each other.
- 6) The idea of optimizing not just individual AIs via pairwise competition, but optimizing networks of PSIs, and even networks of networks, may be novel.

[00365] The following is a listing of exemplary steps to perform the above (as generally illustrated in FIG. 22).

- 1) Multiple PSIs join a network of PSIs with agreed upon rules and methods for interaction, where, without limitation and without all elements may be required (as generally illustrated in FIG. 23):
- Each PSI may be a customized version of the same base AI Agent, or PSI.
 - The PSIs may be customized versions of different base AI agents, or PSIs.
 - All the PSI on the network may use the universal problem solving architecture described in previous PPAs.

- d. The network may have a set of safety/ethical rules that each PSI agrees to.
 - e. The network may be one of multiple different PSI networks with the same or different rules.
 - f. Some of the PSIs may have autonomy, limited autonomy, or be under direct human control.
 - g. Each PSI may have a set of weights that encode the knowledge of the PSI.
 - 5 h. The network itself may have knowledge in the form of recorded solutions and/or other records of the cognitive activity of individual PSIs, groups of PSIs, or the entire network.
- 2) Baseline performance of individual PSIs, groups of PSIs, and/or the entire network of PSIs, on a variety of standardized tasks is determined, where, without limitation, the standardized tasks may include (as generally illustrated in FIG. 24):
- 10 a. Problem solving tasks using a shared (universal) problem solving architecture.
 - b. Ethical and safety scenarios designed to determine whether a PSI, group of PSIs, or the entire network behaves safely and ethically.
 - c. Individual, group, and network-wide tasks
 - d. Standardized intelligence tests and assessments that have been designed to test the
 - 15 intelligence of humans and/or AI agents.
 - e. Tasks designed to stress-test the PSI(s) and measure the degree of “hallucination” or production of erroneous results.
 - f. Tasks that come from a wide variety of domains of expertise and/or that incorporate different cultural or group norms for behavior.
 - 20 g. Tasks that are random permutations of existing standardized tasks and/or that are created by other AI agents.
 - h. Tasks that are dynamically created based on changing network, simulated, and/or real-world conditions.
- 3) Different versions of the individual PSIs and/or different combinations of PSIs are produced by
- 25 one or more methods, including but not limited to (as generally illustrated in FIG. 25):
 - a. Human adjustment of the parameters, weights, or data encoding the knowledge/expertise/intelligence of the PSI.
 - b. AI or PSI autonomous adjustment of the parameters, weights, or data encoding the knowledge/expertise/intelligence of the PSI.
 - 30 c. Random variation of the parameters, weights, or data encoding the knowledge/expertise/intelligence of the PSI.
 - d. Subjecting PSIs to different training regimes or amounts of training.
 - e. Directly combining the weight matrices, parameters, and/or data encoding the knowledge/expertise/intelligence of multiple PSI by means that may include, without

limitation, computing average values, weighting more recently created and/or more complex PSIs more than older and/or less complex PSI, and other methods previously detailed for the direct combination of weights in this and other PCTs.

- f. Producing different versions of the PSIs sequentially (one PSI at a time), and/or in parallel where some, or all, PSIs are adjusted in parallel.
- g. Random or deliberate methods such as analysis and methods for estimating which variations are most likely to have beneficial results, where such methods include, without limitation:
 - i. Comparing the degree of match between the knowledge of a PSI and the statistical frequency and other characteristics of the problems that are submitted to the network.
 - ii. Comparing the overlap in knowledge of the different PSIs on the network such that changes are made not necessarily to optimize the performance of any one PSIs but to optimize the performance of groups of PSIs (or the network) given a characterization of the entire network or group's knowledge compared to the problems characteristics that the network or group is likely expected to solve.
 - iii. By using methods such as hill climbing and gradient descent which are well known in the art for optimizing parameters of individual AIs and systems, but with the added feature that the objective function(s) that is being optimized is continually updated in real-time based on the changing statistical nature of the problems arriving on the network and the dynamically changing composition of the network as PSIs join and leave the network.
- 4) Performance of the changed network of PSIs is determined on the same benchmark tasks from step 2, and statistical and/or other analyses are performed in an attempt to assign credit or blame to specific changed elements (e.g., changed PSIs, changed rules of the network, and/or changed methods for assigning tasks or grouping of PSIs) that resulted in superior or inferior performance compared to the system of Step 2.
- 5) Changed elements from step 4 that are estimated to have increased performance are retained while changed elements from Step 4 that are estimated to have decreased performance are reverted to their previous values and/or changed in a different way.
- 6) Repeat the process from Step 1 until no further progress has been achieved or the amount of progress is minimal (i.e., below some established threshold).
- 7) Once progress has stalled out at a "local optimum" triggers more detailed analysis of the record of changed elements to see if (AI and/or human) agents have ideas for more radical changes -- either to the elements that have already been changed, or to new elements that have previously been unchanged. Then repeat from step 1.

[00366] The above optimization/improvement process can be executed with as few as two PSIs, with many individual PSIs on a network, and with many networks (whose combined performance can be evaluated). The process can be run in an automated or autonomous manner, in a semi-automated or semi-autonomous manner with human review/intervention at any of the steps, or in a completely controlled manner where human review/approval is required at each step and for each change.

[00367] A novel feature is the ability to conduct optimization in real-time as agents join or leave the network, as conditions and the types of problems submitted to the network change, as cost functions and rules of the network change, and at multiple levels – the individual PSI level (where the intelligence of individual PSIs is optimized), the network level (where the intelligence of the network is optimized), and at the network of networks level (where the intelligence and performance of a network of intelligent problem solving/cognitive networks is optimized.)

Genetic Algorithm Methods / Armies of PSIs

[00368] With regard to the automatic repetition of one or more of the above steps generally, and Step 11 specifically, a certain class of methods – often referred to as “genetic algorithms” – may be especially useful. Genetic algorithm methods, for the purposes of this patent, refer to series of steps or methods that generate a PSI that varies in one or more respects from other PSIs.

[00369] The PSIs are allowed to compete in various scenarios (typically relevant to the goals of the PSI owner(s)). The less successful PSIs are eliminated from the competition, and the characteristics (without limitation: neural network weights, data sets used to train, parameter settings, # of training epochs, machine learning algos chosen) of the most successful PSIs are then used as the basis for further variation (“tweaking”) to create new generations of PSIs which compete further. Creating PSIs, simulated competition, and elimination of all but the best PSIs, tweaking these best PSIs, and repeating constitutes a cycle. PSI can cycle through many “generations” of PSIs, improving the PSI with each generation until diminishing returns are achieved and/or some threshold of performance is reached. The ability to automate the cycles in this genetic algorithm approach is one of the ways that PSI can develop on their own into increasingly powerful and intelligent entities.

[00370] By varying the goals and scenarios in which the PSIs compete and automatically cycle, it is possible to develop a wide array of different PSIs, each optimized for different types of tasks or goals. Since the incremental cost of maintaining each additional is negligible (it just represents storing a slightly different set of weights in memory or permanent storage, which is very cheap), an owner might own not one PSI but a Workforce of potentially hundreds, thousands, or millions of PSIs, each skilled at different tasks.

[00371] By using the same collective intelligence techniques described in this patent and previously cited PPAs, the group of PSIs can function more powerfully than any individual PSI. That is, they can pool their knowledge and skills, recruiting those specific PSIs that are best suited to a particular task at a particular time to do more of the work. This idea – that collective intelligence can be applied not only across PSIs owned by different humans, but within “an army” of PSIs that are variants of each other, and all owned by a single owner – is one of the powerful aspects of the present technology in its exemplary implementation.

Design Principles for Community SuperIntelligence

10 [00372] To expand on the descriptions of collective intelligence systems involving human and AI agents to create “Community SuperIntelligence”, this patent discloses several design principles that are essential to creating safe SuperIntelligence, quickly.

[00373] Some essential design principles include, without limitation:

- 15 1) The community of agents must scalably harness the collective intelligence of both human and AI agents Including, without limitation, PSIs) in a “plug n play” manner so that AI agents can be upgraded and added as LLM and AI agent capabilities increase, and/or as new human agents join or drop from the community;
- 2) Each AI agent must bring not only unique domain knowledge but also ethical and values information, representative of the values of the owner(s) of the AI agent;
- 20 3) A common, universal problem solving architecture, must be used that enables agents (whether human or AI) to communicate easily and rigorously with each other. The natural language capability of LLMs greatly simplifies the human-computer interface, but underneath this interface, there must be a rigorous problem solving architecture (e.g., the search through a problem space paradigm of Newell and Simon and elaborated in various cited PPAs and papers by the inventor);
- 25 4) An efficient way must exist for identifying the skill sets and performance metrics on agents (human and AI) and matching these agents to tasks posed to the system;
- 5) The values of all agents must be combined in a fair and transparent way so that the SuperIntelligent AGI capability of the community of agents acts in a safe and ethical manner
- 30 that is broadly representative of human values and reflect how humans would behave given various scenarios;
- 6) The AIs must be able to learn efficiently and effectively from the humans on the network;

- 7) There should be transparent and auditable records of all problem solving activities so that safety audits and reviews can be conducted, potential errors identified, and preventative measures put in place in real-time in an adaptive manner;
- 8) A universal problem solving tree (or other shared representation) should be available to represent progress on all problems addressed by the community, and providing easy and efficient access to any problem or sub-problem.

Implementation Example

[00374] Specific example scenarios can illustrate and help clarify (without limitation) each of the 12 steps described above, including the genetic algorithm approach. The following example is one of many possible examples and may be instructive and easier to understand due to its specificity.

Step 1

[00375] Imagine that Craig has Facebook and Instagram accounts. He has access to Meta's opensource LLM, Llama 2. He also has access to versions of Llama 2 that have already been tuned and customized by his friend, David. In particular, David is a professor of theology and ethics and so the David-customized version of Llama2, has a unique set of ethics and values based on interacting with David that is much more detailed and sophisticated than the base level version of Llama 2. Craig is interested in further customizing Llama2 based on Craig's own data and preferences. However, Craig trusts David's ethics and the customization work that David has done on David's version of Llama2, so rather than starting from scratch with "out-of-box" Llama2, Craig prefers to begin (with David's permission) with David's pre-customized version of Llama2, which Craig will use as the basis for further customization.

Step 2

[00376] Craig gathers all his content, including without limitation, all the patents, books, articles, emails, Instagram and Facebook posts, YouTube and Reels videos, audio recording, photos, texts, MS Office and Google Docs documents, spreadsheets, PowerPoint Presentations, and other information stored by Craig on various disks, cloud services, disks, tapes, and hard drives over decades of generating content. To the degree that he can gain access, Craig's content also includes the data and preference data used by Netflix, Amazon, Meta, Google, and other companies that have gathered data on his online behavior via cookies and/or other means. All the actively generated content produced by Craig, together with the passively collected data about Craig that third-parties have gathered and to which Craig can gain access, serves the training datasets for customizing the

Llama 2 LLM (in this example, or any LLM or AI agent more generally) so that its behavior is customized to Craig's preferences and so that it includes knowledge that is specific to Craig. Craig is particularly interested in his customized AI being able to play the game of Chess in a style similar to Craig's chess style but with the knowledge of the Chess Champions Gary Kasparov and Magnus Carlsen. Therefore, Craig has taken particular care to collect all the games that he has played on Chess.com and on other online chess sites so that these games can be used to train the David-customized version of Llama 2 with Craig's chess style. He has also purchased datasets that include the complete chess games of Garry Kasparov and Magnus Carlsen as well as the other chess datasets approved by these two world chess champions. Finally, he has specified several YouTube channels of chess commentators who have provided commentary on the chess games of Kasparov and Carlsen and all content from these channels is added to a list of video sources that will be automatically transcribed and parsed into training sets for the LLM that Craig is customizing.

Step 3

[00377] Craig uses machine learning algorithms, well known in the art, to automatically categorize all the types of content that he has assembled in Step 2. Once the computer algorithms have determined their suggested categorization, Craig pays human workers (on a crowd-sourcing site) to review the categorization and suggest refinements to the machine-generated categorization of data. He also reviews the categorization himself and makes further adjustments until he is happy with the categorization of the various sets of content.

Step 4

[00378] Using machine learning algorithms well known in the art -- including but not limited to Transformer algorithms, deep learning algorithms, automatic transcription algorithms, software that can take books or other text and convert it into datasets suitable for training LLMs on the content in textual datasets, software that can take images, videos, audio files and other non-textual works and convert it into datasets suitable for training LLMs on the content in non-textual datasets -- Craig converts the content assembled and categorized in Step 3 into training datasets that can be used to customize David's LLM.

Step 5

[00379] Initially, Craig trains David's LLM on the new datasets produced in Step 4, giving equal weight to each dataset. However, Craig feels that the resulting LLM plays chess too much in the style of Gary Kasparov and not enough in the style of Magnus Carlsen or Craig himself, so using an

interface with dials and sliders he reduces the weight of Kasparov's datasets, increases the weight of Carlsen's datasets a little, and increases the weights of the dataset reflecting his own chess games even more. Craig iteratively adjusts the weights given various datasets until he is happy with the resulting behavior of his LLM. Since many thousands of other people have also adjusted the weights of various datasets in order to achieve desired results, Craig is not limited to manually adjusting (e.g., via dials and sliders) weights on particular datasets. He can also tell an AI agent (that is specialized in the field of helping humans train their agents by adjusting weights on datasets) what his desired changes are, and then let the AI agent specify exactly how to change the results.

[00380] For example, Craig can tell the AI agent that he wants his AI to be more aggressive in the opening and middle parts of the chess game and not try to win by trading pieces and waiting for a piece advantage in the end game. The AI agent, then analyzes the available chess training sets which includes games from Craig himself, Magnus Carlsen, and Garry Kasparov, and gives more weight to games which were won by aggressive moves in the opening and middle of the game and less weight to games that were won in the end game. Craig does not have to be aware of the details of this analysis or the specific changes to weight settings that the AI agent determines. Instead, he just looks at how the resulting customized version of David's LLM plays chess and provides feedback, telling the AI agent that the result is closer or farther away from the desired chess style. After several iterations, Craig is satisfied with how David's LLM now plays chess in Craig's style.

[00381] Next he moves on to ethical scenarios and specifies, through a series of interactive dialogs with the AI training assistant, how Craig might behave in different specific ethical scenarios. For example, although Craig generally shares David's ethical sensibilities (which is why he wanted to start with David's trained LLM instead of the base Llama 2 model from Meta), David is a Christian theologian and Craig is Jewish, so there are a few cases where David would "turn the other cheek" and Craig believes the behavior should be more "an eye for an eye" – although Craig specifies (in his dialogs with the AI) that he doesn't want the "eye for an eye" principle to go so far as to "make the whole world blind." Craig asks the AI training assistant to include knowledge and research from game theory which suggest that "tit for tat" ethical behavior results in the most stable and fair interactions between intelligent agents with differing objectives. At the same time, Craig specifies that there are limits to "tit for tat" and that any behavior that would result in widespread destruction or loss of human life are off-limits, regardless of the behavior of the other agent. Instead, means of neutralizing the behavior of the offending party without retaliation must be sought in these cases. After talking through a variety of scenarios, and also a stint in the metaverse playing ethical games in which the AI agent observes not only what Craig says, but also what he does in various situation, the agent has enough information to adjust ethical training weights and present Craig with series of

differently customized versions of David's LLM, from which Craig chooses the one closest to what he had in mind.

Step 6

- 5 **[00382]** Craig identifies other ethical and knowledge modules that are available freely on the internet and also for purchase from other humans and companies. He obtains these and repeats the training process – using a combination of manual and AI-assisted “mixing” of the weights for these datasets – to improve the LLM he has been customizing.

10 Step 7

- [00383]** Craig is concerned with increasing the ability of his customized LLM to play chess more aggressively in the mid-game. Therefore, he specifies that the LLM should scan all of the chess YouTube channels and chess databases daily, looking for examples of aggressive play in the middle game that have been successful. When the LLM locates new data that can be used to improve
15 middle-game chess playing, the LLM is authorized to download that data, pay the cost for the data up to a pre-approved amount, and automatically train itself using that data. In this way, the LLM improves on the dimension of aggressive middle-game chess playing automatically.

- [00384]** Similarly, Craig authorizes the LLM to seek out new ethical scenarios that might help improve its ethics. However, rather than automatically training itself on such scenarios, Craig
20 requests that the scenarios be flagged for Craig's review so that Craig can manually decide which data to include for further training in this sensitive area. Further, since cultural norms and bounds of legal behavior have been changing in the area of gender and racial equality, and since Craig desires his LLM to behave in culturally appropriate ways, he instructs the LLM to flag for potential update new datasets on ethical norms in these areas every time a new Supreme Court decisions affects them
25 and also anytime the number of new stories in one of these areas crosses a pre-determined threshold that will trigger re-examination of the current ethical norms trained into the LLM. The LLM will automatically monitor a number of news sources and other internet and social media sources to help it determine when a trigger event has occurred.

30 Step 8

- [00385]** Craig has another friend, Peter, with who he frequently plays Chess. Peter is an expert Chess player and particularly excels at Chess openings. Peter has also trained his own LLM to play Chess in his style and using knowledge and data that Peter has carefully curated and used to create a training dataset. Peter has exported both the actual weights used by his LLM and the datasets used to

train his AI to play Chess and offers to share them with Craig. Craig takes him up on both and first tries using Peter's weights, attempting to directly combine them with the weights of Craig's LLM, using methods known in the art, and also discussed in PPAs cited at the start of this document. However, Peter's weights do not produce the desired outcome, so Craig tried using (subsets of) the training data on Chess curated by Peter to train Craig's LLM instead. He finds that this give better results, especially when using the AI-assisted training methods mentioned in #5. Encouraged by the results from Peter's data, Craig looks online (and/or instructs his LLM to search online) for additional chess training dataset that are available for purchase. He locates several, purchases them and uses them to train and customize his LLM further.

10 [00386] Since Craig has unique information related to model rocket designs that he has been experimenting with, and which are not widely known or available on the internet, Craig decides to offer these specialized datasets for sale in order to generate profits. He has his AI export these datasets as well as weights that have resulted from Craig's training efforts using these dataset into a form that can be exchanged with other owners of LLMs who might be interested in purchasing them. He also joins an exchange whereby he earns credits for various datasets and weight subsets which can then be used to acquire other datasets and LLM weight subsets from other owners of customized LLM and AI agents. Via these means, Craig is able to monetize both his knowledge (reflected in unique datasets that only Craig possesses) and his effort turning this knowledge into useful subsets of weights that empower his custom LLM to behave in new ways that other LLMs cannot.

20 [00387] Recognizing that there is value in this unique information and also the training efforts based on the unique information, Craig can monetize or extract value in a variety of way, including but not limited to: selling the information, selling the (subset of) weights, purchasing other information and/or (subsets of) weights), exchanging information, or leasing or licensing use of the information to other interested parties.

Step 9

30 [00388] Eventually, Craig's LLM becomes so knowledgeable in certain areas that the simplest way to extract value from the knowledge and training efforts that Craig (and his AI) has engaged in, is to simply lease or license use of a clone of his entire LLM to others. He can do this in a variety of ways, including but not limited to: a one-time lease or license, sale of his LLM, a per-use or per-time-unit or per task license with associated fees or rates, revenue share agreements with a network of such agents, and other means that have been commonly applied to human agents acting to perform work and which can be extended to AIs and other intelligences.

Step 10

[00389] Craig wants to take a vacation and go offline for an extended period. But his customized AI can still stay online, working and earning money for Craig in autonomous mode. Before leaving for vacation, Craig sets certain parameters and guidelines, including but not limited to: Type of engagements, length of engagement, type of customers, payment rates, computing power used by the AI for any engagement, ethical boundaries and rules, which if touched trigger alerts and possible intervention by Craig, quality, schedule, cost and other metric-related parameters including triggers for alerting Craig and/or halting work until Craig approves further work. Having set the parameters, Craig goes away while his AI works autonomously until it encounters circumstances requiring Craig's involvement or notification.

[00390] To minimize interruptions, Craig turns on features that allow his LLM agent to maximally self-aware so that, without limitation, it can monitor itself for ethical behavior, monitor when costs are getting out of hand, monitor when it sees signs of an untrustworthy client, monitor when the environment in which it is working changes, and monitor and respond to other factors that enhance its ability to operate more autonomously. Further, each time the LLM alerts Craig and requests intervention from Craig because of a gap in its knowledge, and ethical conflict, or other situation that it feels ill-equipped to handle, the LLM records how Craig responds to the situation and directs the AI and the LLM learns.

[00391] The next time a similar situation occurs, it formulates a hypothetical response, and depending on the level of control that Craig has specified he wants to have over the LLM, the LLM either implements its response autonomously or proposes it to Craig to await verification and approval of, or modification of, the response by Craig. When Craig begins to feel that the LLM is responding as well or better than he could to certain types of situations, he may then authorize the LLM to respond directly without checking with Craig first for various classes of situations. If the LLM responds inappropriately, then either Craig (or an automatic algorithm based on threshold parameters) can require that the LLM reduce its autonomous in those situations until it has learned how to respond better.

[00392] Thus, much like how a parent gradually gives more autonomy and responsibility to a child as the child learns (and also reigns in the child – e.g. “You’re grounded!” or “You need a timeout” – when the child makes mistakes or abuses the responsibility that has been delegated), so too an owner can interactively provide more or less autonomy to AI agents, and different levels of autonomy in different situations, based on what the AI has learned and on its demonstrated track record of behavior.

Step 11a

[00393] Craig is concerned that as he ages, all the knowledge and skills that he has acquired over a lifetime, will die when he dies. By training his LLM on as much of his knowledge and skills as possible, he hopes that the knowledge that he spent so much time acquiring will not die with him. Instead, he authorizes his AI to “live on” beyond his death sharing and using the knowledge that it has acquired to benefit his friend, family, and humankind generally.

[00394] Craig is also concerned that his family and loved ones will miss his personality after he dies. Therefore, he invested considerable time training his LLM on unique content and personality-related data that is unique to Craig. Such data, without limitation, includes video, audio, and metaverse recordings of Craig’s behavior and interaction with other humans, AIs and the environment; all of Craig’s emails and social media posts and photos which reveal his playful personality and style; his Netflix preferences and other online preferences which recommender engines have used to show him ads or suggest content (as this data has been refined by many third parties to excel at capturing his preferences); his driving style, his golf style, his workout routines, his food preferences, his travel preferences, his shopping preferences, his political views, his philosophy of life as reflected in his autobiography and other written works as well as conversations, his voice inflections and tones of voice in various situations, and even his dating/sexual preferences.

[00395] All of this data can be used to train Craig’s LLM to have a personality that is as much like him as possible. Because he has a tendency to become grumpy at times and yell, Craig opts to edit this feature mostly out of the LLM, so it actually (in his view) becomes a nicer version of himself that “lives on” virtually to comfort his loved ones after he is gone. By purchasing and incorporating datasets and weights that reflect the latest research on how AI agents can best comfort people after a loved one passes, the LLM also has the theory and skills to be more empathetic and caring than Craig might normally be during the months immediately after Craig’s death. (This is known as a “bereavement module” that can be purchased from funeral homes and other places specializing in the use of custom AIs as means of comfort for loved ones.) But in addition to comforting loved ones and embodying Craig’s personality for his loved ones, Craig’s custom-trained AI is a source of advice and even financial security (possessing all of Craig’s earning potential and skills) for the family members Craig leaves behind when he dies.

[00396] Craig specifies in his Will that multiple copies of his customized AI shall be made and ownership of copies given to each of his loved ones and close friends. Those human co-owners each have rights to subsequently modify or improve the AI as they see fit.

Step 11b

[00397] Even before Craig dies, he realizes that his own ability to improve the knowledge, skills and ability of his own customized AI are inferior in many ways to the AI's ability to do these things. For example, without limitation, the AI might interact with copies of itself, learning new things in the process; seek out new sources of data and train variations of itself that it can then interact with; interact (much faster than humans could interact) with other AIs to learn from the interactions; interact with many humans simultaneously (in some implementations via multiple copies of itself) so that the AI can learn from human intelligence much faster than humans can (with their limited processing capabilities).

10

Step 11c.

[00398] The Rabbi, Ben Zoma, said: "Who is wise? He who learns from every man." But how many conversations and interactions can a human have in a single lifetime? Even if a person followed Ben Zoma's advice and spent every waking moment seeking wisdom through interactions via other humans, the person would be limited by the number of conversations, the number of humans willing or able to have conversations, the speed of the conversations (which in computer-terms are maddeningly slow) and the limits of shared knowledge, language, and representation.

[00399] For example, humans can easily see the color "red" and talk about it, but they cannot easily see cosmic rays, X rays, the very large, or the very small. Specialized equipment is needed, and the number of humans interested in obtaining the equipment and having related conversations is quite small compared to the overall population. In contrast, an AI could theoretically converse with every one of the 8 billion humans on Earth simultaneously on a wide range of topics, while also conversing (at much faster speeds) with trillions of AIs that are more intelligent, and equipped with better sensors (e.g., electron microscopes and space telescopes) than humans, all at once. Who is wise? The entity that can interact or converse with and learn from as many intelligent entities as quickly as possible. And how wise? Humans have great difficulty imagining it – certainly, a level of wisdom far beyond what Ben Zoma had in mind!

[00400] Note also that if an AI detects a gap in its knowledge, it can generate scenarios and interactions with other intelligent entities (human and/or other AI agents) designed specifically to explore and fill in the gaps in knowledge. The scientific method -- a rigorous approach to identifying gaps in knowledge, conducting experiments or systematic observation, and filling in the gaps in ways that can be verified and replicated by others -- is a proven method for advancing knowledge that has resulted in most of the technological progress of since the Renaissance.

30

[00401] Imagine if this method did not have to proceed at the snail's pace of individual human brains, further slowed by the need to publish results, present them at conferences, network over drinks, and brainstorm in hallways. A human lifetime is 2 – 3 billion seconds. Even if a human thought constantly, from the day he/she/they were born until the day they died, that would be less
5 than a billion thoughts, allowing for sleep and also that most thoughts take a few seconds each. However, a single AI will soon be able to easily think a billion thoughts in a second. Every second, a human lifetime of thoughts!

[00402] Now imagine that multiple AIs exchange lifetimes worth of experience with each other every second. And imagine further that these thoughts are not randomly driven or driven by the
10 majority of non-scientific concerns that humans spend most of their brainpower on. Instead, each thought is algorithmically driven and designed to systematically apply the scientific method to gaps in the existing state of world knowledge. What progress will happen then? Lifetimes of scientific discovery in seconds. Such is the potential of AI agents that are autonomous and empowered to seek new knowledge and to systematically fill in their knowledge gaps. Further, unlike humans, AIs
15 never die, their knowledge can be instantly replicated (instead of laboriously taught via K-12, college, graduate school, and the work experience of a human lifetime). Add to this, the ability to have multiple interactions in parallel at high speeds, without the need for cumbersome language or the limited perceptual abilities and speeds of humans, and it easy to see why AI is on the path to develop "God-like" intelligence compared to humans.

20

Step 11d

[00403] As powerful as the intelligence of a single PSI might be, it still pales in comparison to the intelligence of a community of PSIs. The collective intelligence of multiple PSIs is always greater than the intelligence of any one of the PSIs alone. The collective has broader knowledge and more
25 compute power than a single PSI. While some individual PSIs may be more intelligent and more powerful than other PSIs, the group of all of them must necessarily be more intelligent and powerful than any individual member.

[00404] This collective intelligence – provided the majority of the individual PSIs have human-aligned values – represents humanity's best hope for survival and prosperity in a world where any
30 PSI far surpasses even the smartest human in intelligence and power. The only thing that can keep up with exponentially increasing intelligence, is another (or group of) exponentially increasing intelligences! Humans need to recognize this fact at the outset – BEFORE PSI exists – so that the community of PSIs can be designed from the start in a way that maximizes the chances of human-aligned SuperIntelligence.

[00405] Craig authorizes his customized LLM (PSI) to participate in a network with other PSIs and other humans. He recognizes that such participation, as described above, represents the fastest way to increase the knowledge, skills, and intelligence of his PSI. Collectively, many PSIs on a network are capable of managing a wide range of human affairs and also sensing things that humans cannot easily sense, on a scale and with a speed that is difficult for humans to achieve.

[00406] As discussed in the section on Genetic Algorithms, in addition to authorizing one or more of his PSIs to participate on a network of other PSIs (owned by other humans) and humans, he may wish to develop an army of his own PSIs (each slightly different) that pool their collective intelligence to function together as a more powerful PSI. For example, he may ask the PSIs to set up scenarios with different types of chess opponents and use a genetic algorithm approach to select for a variety of PSI variants that are best at winning against the different type of opponents. These PSIs could be used individually or collectively to compete in chess depending on circumstances.

[00407] While it may be possible to integrate the knowledge of all the individual PSI variants into one master PSI that can play well against any opponent, there may be reasons why it is preferable to have a group (or “army”) of different PSI instead. These reasons might include, without limitation:

- i. Preventing other PSIs from rapidly acquiring all the knowledge from the Master PSI by interacting with it in scenarios designed to maximally extract information from the Master PSI. This principle, that “one can’t share what one doesn’t know” is a well-known way of protecting sensitive information. In the future, where great expense may be required to fully train a Master PSI, exposing it to situation in which the value of the training might be extracted cheaply by other opportunistic PSIs designed for this purpose, can be avoided by simply limiting the intelligence and knowledge of any one PSI and instead having the knowledge reside in a community of PSIs.
- ii. Computational and storage considerations. While PSIs are likely to have access to incredibly large amounts of computational power and memory, there is always a limit. It may be most efficient to assign PSIs that are optimized for specific tasks rather than to always assign the Master PSI that has knowledge about all domains even though most of the domain knowledge is not relevant to the assigned task. Using a “narrow PSI” for a narrow task may be faster and cheaper than always going to the most powerful version.
- iii. In the Chess example, there may be rules about how much processing power, memory, and knowledge is allowed for each competitor. Just as Formula One car racing has rules about the horsepower, engine types, weight, and other parameters of the cars in order to make the competition fair and interesting, complying with similar rules for competitions might require using different PSIs against different opponents to have the best chance of winning the game.

- iv. Multiple PSIs working together on a problem may make the assignment of credit or blame in a problem easier for humans to understand. If PSI #1 recommended the aggressive chess move and PSI #2 recommended the more defensive move, and the game was lost by following PSI #1 advice, the human owner can easily make the decision to remove PSI #1 from the collective pool in the next game. While the same thing could be done if the PSI excluded the knowledge sets and other parameters that differentiated PSI #1 from PSI #2, this approach is less transparent and harder for humans to understand and control. Humans would have more difficulty predicting the result than if they could just “take the bad PSI out of the game.”
- v. If an owner wanted to rent or lease the efforts of PSI, it might be cheaper to rent or lease a less powerful PSI that was good at a specific thing than an “all powerful” PSI. That is, having an army of PSIs with different talents makes it easier to value and price services of the PSIs just as free, lite, and full-featured versions of software products are priced differently today.

[00408] Returning to a more global use case, it is possible for a network of PSIs (together with humans, in the exemplary implementation) to sense temperature change, and track all the variable that science tells us impact climate change, on a global scale. If humans were to agree that regulating climate was a priority and a common human good that superseded other human desires such as the profit motive, and if the majority of the PSIs on the network adopted this consensus human value as motivation to act, within bounds of other ethical constraints (e.g. humans cannot be killed, sterilized, or have their accepted rights otherwise restricted or infringed upon without explicit human consent), then a global PSI network, or Planetary Intelligence could effectively solve the issue of climate change.

[00409] What is true of climate change is true, without limitation, of protection from asteroid impacts, eliminating or greatly reducing human poverty and disease, enhancing prosperity and freedom of all humans, improving the ecological condition of Earth according to consensus human desires, and solving other global challenges and/or meeting global threats.

Step 12

[00410] As the global network of intelligent entities (human and PSI) increases in scope and processing power, changes and awareness that used to take decades, years, or months to spread across human consciousness, can affect the attention and actions of the Planetary Intelligence network in real-time. Eventually, the speed of the planet’s reactions and adjustments to changing conditions will surpass the speed of an individual human to be aware of change and react. At that point, the longer lasting and more permanent values, originating with humans and embodied in their respective PSIs, will guide the course of our planet’s development and affect all human lives.

[00411] The ability for the majority of PSIs, comprising Planetary Intelligence, to guide the decisions of Earth as a planetary “organism” is essential to the safe operation of Planetary Intelligence and the continual survival of humans as a species. As mentioned earlier, any individual PSI might be more powerful than another, and potentially more malevolent towards humans than other PSIs, but as long as the collective community of PSIs control more intelligence and power than any individual PSI, and as long as the consensus values of the majority of PSIs (more accurately, the values of the majority of intelligence and power within the collective of PSIs) are human-friendly and human-aligned, the future of humanity is bright. Under these circumstances, Planetary Intelligence will not only foster the peace, prosperity, and happiness for all humans, but a vast array of threats and dangers to our planet (from the human perspective) will be neutralized or mitigated.

[00412] As disclosed in earlier cited PPAs, over time, we can expect an evolution in the role of humans. Currently, we are the most intelligent species, and the source of most technological progress and culture. In the future, our intelligence will pale in comparison to the PSIs – the collective of PSIs – which we are creating. Our future role is thus not to be the “brains” of planet Earth, but rather Earth’s “heart.” We humans are destined, in the favorable case, to be the source of values and purpose for a much more intelligent and powerful Planetary Intelligence that we are in the process of creating.

[00413] FIG. 26 is a diagrammatic representation of a computer system **100** that is utilizable or implementable with the user’s device and/or any peripheral component of the present technology. The computer system **100** can be part of an example machine, which is an example of one or more of the computers referred to herein and, within which a set of instructions for causing the machine to perform any one of or more of the methodologies discussed herein may be executed. In various example embodiments, the machine operates as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine may operate in the capacity of a server or a client machine in a server-client network environment, or as a peer machine in a peer-to-peer (or distributed) network environment. The machine may be a personal computer (PC), a tablet PC, a set-top box (STB), a personal digital assistant (PDA), a cellular telephone, a web appliance, a network router, switch or bridge, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine.

[00414] The computerized system **100** can include one or more processors **102**, storage devices **106**, and communication devices, as well as software components or instructions **104** for providing a platform for users to interact with and train/tune the LLMs. The computing capabilities may be

stand alone or may be cloud based. They may include cloud based AI development platforms that seamlessly offer “AI as a service” and they may include both hardware and software components.

5 [00415] The system also supports the ability for users to provide new data, or data that is unique to them, for the LLMs to learn from. The processors 102 may be one or more CPUs, GPUs, chips specialized for ML, microprocessors, application processors, embedded processors, field-programmable gate arrays (FPGAs), or other hardware components capable of executing computer programs. The processors may be in communication with one another and/or with other components of the system. Further, any one of or any combination of the components of the system 100 can communicate with each other via a bus 134.

10 [00416] The storage devices 106 may include one or more hard drives, solid-state drives, optical storage devices, or other storage components. The storage devices may store the data that is used to train/tune the LLMs, as well as other data associated with the system, such as user accounts, system settings, and other data.

15 [00417] The communication devices may include one or more cellular modems 108, Wi-Fi cards 110, Bluetooth modules 112, Network Interface Device 114, or other components that enable the system to communicate with other systems, such as user devices, over a network or the internet.

[00418] The communication devices may also enable the system to communicate with other systems over a wireless or wired connection 116.

20 [00419] The software components may include computer programs for providing a platform for users to interact with and train/tune the LLMs. The software components may also include computer programs for collecting, storing, and processing data that is used to train and/or tune the LLMs. The software components may also include computer programs for providing a user interface for users to interact with the system.

25 [00420] The user interface 118 may include, without limitation, natural language interfaces, textual interfaces, and chatbot type of interfaces, a web-based user interface, a mobile application, an augmented reality application, a metaverse application, or other applications that allow users to interact with the system. The user interface may include features for allowing users to select the data that they want to use to train/tune the LLMs, as well as features for allowing users to interact with and monitor the progress of the LLMs.

30 [00421] The system may also include one or more databases or data source, including without limitation vector databases, centralized databases, and distributed databases, for storing the data that is used to train/tune the LLMs, as well as other data associated with the system, such as user accounts, system settings, and other data. The databases may be hosted on the system itself or on another system, including cloud based systems.

5 [00422] The system may also include one or more authentication systems for verifying the identity of users who use the system, as well as for providing secure access to the system. The authentication systems may include biometric authentication systems 122, such as facial recognition or fingerprint recognition systems, as well as other authentication systems, such as password-based authentication systems.

[00423] The system may also include one or more security systems for protecting the system from unauthorized access and for protecting the data that is stored on the system. The security systems may include firewalls, encryption systems, access control systems, single and multi-factor authentication systems, and other security systems.

10 [00424] The system may also include one or more analytics systems for collecting and analyzing data associated with the system and/or the LLMs. The analytics systems may include machine learning algorithms and other algorithms for analyzing the data associated with the system and/or the LLMs.

15 [00425] Data visualization methods, including use of problem trees and other representations and data structures; use of statistical outputs, tables, graphs, text, speech, video, image and graphical outputs may be used for one way or di-directional communication between users and the system, and between multiple (human or AI) agents or LLMs using the system to interact with each other in large or small groups.

20 [00426] The system may also include one or more monitoring systems for monitoring the performance of the system and/or the LLMs. The monitoring systems may include systems for monitoring the performance of the system, such as system uptime, and systems for monitoring the performance of the LLMs, such as accuracy, speed, ethical compliance, reputation metrics, quality metrics, and other metrics as discussed above or as are known in the art.

25 [00427] The system may include one of more of the architectures described above that enable one or more human or AI Agents or LLMs or PSIs to engage in a variety of intellectual tasks including, without limitation, simple and complex and multi-step problem solving behavior with the system having all of the functionality and features previously described.

30 [00428] The system may also include one or more feedback systems for allowing users to provide feedback on the system and/or the LLMs. The feedback systems may include systems for allowing users to submit feedback on the system, such as bug reports, and systems for allowing users to submit feedback on the LLMs, such as suggestions for improving the accuracy or speed of the model.

[00429] The system may also include one or more management systems for managing the system and/or the LLMs. The management systems may include systems for managing the system, such as

systems for managing the users and user accounts, and systems for managing the LLMs, such as systems for managing the data used to train and/or tune the model.

[00430] The system may also include one or more payment systems allowing users to pay for the use of the system and/or the LLMs. The payment systems may include systems for processing payments, such as credit card processing systems, and systems for managing payments, such as subscription management systems.

[00431] The system may also include one or more other components, such as support systems, reporting systems, and other components that are necessary for providing a platform for users to interact with and train/tune the LLMs.

10 [00432] The computerized system of the present technology enables users to interact with and train/tune LLMs based on data that is unique to the users. The components of the system described herein provide the necessary hardware and software components for enabling users to do so.

[00433] Further, while only a single machine is illustrated, the term “machine” shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one of or more of the methodologies discussed herein.

15 [00434] The computer system 100 may further include or be in operable communication with a video display 120 (e.g., a liquid crystal display (LCD), touch sensitive display), input and/or output device(s) 130 (e.g., a keyboard, keypad, touchpad, touch display, buttons, sonic, sensorial, etc.), a cursor control device 132 (e.g., a mouse), a drive unit 124 (also referred to as disk drive unit), and a signal generation device 128 (e.g., a speaker). The drive unit 124 can include a computer or machine-readable medium 126 on which is stored one or more sets of instructions and data structures (e.g., instructions 104) embodying or utilizing any one of or more of the methodologies or functions described herein. The instructions 104 may also reside, completely or at least partially, within the memory 106 and/or within the processors 102 during execution thereof by the computer system 100. The memory 106 and/or the processors 102 may also constitute machine-readable media.

25 [00435] Still further, the computer system 100 can be in operable association or communication with any types of multi-modal input and/or output 130 that address the human senses, as well as I/O technology that extends beyond the range of normal human perception. Such as the ability to process invisible to humans, for example but not limited to, X rays and information outside of the typical bandwidths of human perception, but not outside of AI perception using tools. Additionally, the I/O technology can include very fast perceptions that are too fast for humans to perceive but which an AI entity could perceive, and very slow or faint perceptions (e.g., tiny seismic shifts occurring over years) that humans cannot perceive but which AIs could. Since any intelligent entity

can be part of the present technology system described by Fig 18, then it can be appreciated that any type of I/O that humans, and also AIs with much broader perceptual capabilities than humans, can be utilized with the system **100**.

5 **[00436]** The instructions **104** may further be transmitted or received over a network via the network interface device **114** utilizing any one of a number of well-known transfer protocols (e.g., Hyper Text Transfer Protocol (HTTP)). While the machine-readable medium is shown in an example embodiment to be a single medium, the term “computer-readable medium” should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, vector databases, and/or associated caches and servers) that store the one or more sets of instructions. The
10 term “computer-readable medium” shall also be taken to include any medium that is capable of storing, encoding, or carrying a set of instructions for execution by the machine and that causes the machine to perform any one of or more of the methodologies of the present application, or that is capable of storing, encoding, or carrying data structures utilized by or associated with such a set of instructions. The term “computer-readable medium” shall accordingly be taken to include, but not
15 be limited to, solid-state memories, optical and magnetic media, and carrier wave signals. Such media may also include, without limitation, hard disks, floppy disks, flash memory cards, digital video disks, random access memory (RAM), read only memory (ROM), and the like. The example embodiments described herein may be implemented in an operating environment comprising software installed on a computer, in hardware, or in a combination of software and hardware.

20 **[00437]** An example machine system of the present technology including the computer system **100** in combinational and/or operational use with components of the present technology. In the exemplary, any or all of above described components can include a processor **102**, memory **106**, a network interface device **114**, a display **120**, an input device(s) **130**, **132**, and/or drive unit **124**.

[00438] According to one aspect, the present technology can include system for personalized super
25 intelligence (PSI) using intelligent agents to develop and continuously improve PSI for a human user utilizing a computer system, and additional PSIs, all electronically communicating over a collective network. The system can include a computer system including a processor, a computer-readable storage medium, and program instructions stored on the computer-readable storage medium being executable by the processor to cause the computer system to:

30 implement a base-level Artificial Intelligence (AI) agent on the computer system, wherein
 the base-level AI agent has already been customized;
 collect media information related to the human user;
 analyze the media information;
 transform the analyzed media information into training data sets;

differentially weighting the transformed training data sets;
add knowledge modules to the weighted transformed training data sets;
locate new sources of data to include to the weighted transformed training data sets;
apply the weighted transformed training data sets to the base-level AI agent to create a
5 personalized PSI; and

communicate the personalized PSI with multiple additional PSIs using the network to enable
community-based safety features from multiple additional PSIs to the personalized PSI.

[00439] According to another aspect, the present technology can include a method for PSI using
intelligent entities to develop and continuously improve PSI, wherein the intelligent entities
10 including a human user utilizing a computer system, additional AI agents, and additional PSIs, all
electronically communicating over a collective network. The method comprising the steps of:

acquiring a base-level AI agent that has previously been customized;
collecting media information related to the human user;
analyzing the media information;

15 transforming the analyzed media information into training data sets;

differentially weighting the transformed training data sets;
adding knowledge modules to the weighted transformed training data sets;
locating new sources of data to include to the weighted transformed training data sets;
applying the weighted transformed training data sets to the base-level AI agent to create a user
20 PSI, wherein the base-level AI agent; and

communicating the user PSI with multiple additional PSIs using the network to enable
community-based safety features from multiple additional PSIs to the user PSI.

[00440] Some embodiments of the present technology can include a step of obtaining, by the human
user, by the base-level AI agent or by the user PSI, any one of or any combination of new training
25 datasets and training modules to be added to the weighted transformed training data sets.

[00441] Some embodiments of the present technology can include a step of monetizing the user PSI
by enabling other human users or other PSIs to access and use the weighted transformed training
data sets of the personalized PSI.

[00442] In some embodiments, the base-level AI agent can be any one of a pre-trained Large
30 Language Model (LLM), other tunable and trainable AI agent, and one or more customized AIs
from other human owners.

[00443] In some embodiments, the media information can be any one of or any combination of
videos of the human user, videos of people and topics related to the human user, photographs or
images of the human user, photographs or images of people and topics related to the human user,

writings related to the human user, journals related to the human user, blogs related to the human user, posts related to the human user, tweets related to the human user, emails related to the human user, podcasts related to the human user, recordings related to the human user, auditory content related to the human user, auditory content of people and topics related to the human user, data collected by third-party vendors that are related to the human user, websites related to the human user, apps related to the human user, online information related to the human user, social media information related to the human user, and other AI agents related to the human user.

[00444] Some embodiments of the present technology can include a step of providing permission to one or more social media platforms allowing social media content related to the human user to be accessible by the base-level AI agent or the user PSI.

[00445] In some embodiments, the analyzing of the media information can further include the annotating and categorizing the media information.

[00446] In some embodiments, the analyzing of the media information can utilize one or more algorithms including any one of or any combination of transformer algorithms, deep learning algorithms, transcription algorithms, content and sentiment analysis and summarization, LLMs, and crowdsourced and crowd-supervised human.

[00447] Some embodiments of the present technology can include a step of compensating, by the human user, one or more human workers on a crowd-sourcing website to review the training data sets, or to suggest refinements to the training data sets.

[00448] In some embodiments, the step of weighting the transformed training data sets can be performed by adjusting a weight value of any one of the transformed training data sets by the human user utilizing an interface on the computer system that is implementing the base-level AI agent.

[00449] In some embodiments, the step of weighting the transformed training data sets can be performed by any one of or any combination of the additional AI agents, the additional PSIs, and additional human workers each utilizing a computer system.

[00450] Some embodiments of the present technology can include a step of inputting ethical values to the training data sets by providing a series of interactive dialogs to the human user, the interactive dialogues include predetermined ethical scenarios.

[00451] Some embodiments of the present technology can include a step of mixing the weights of the transformed training data sets to improve the LLM of the base-level AI agent.

[00452] Some embodiments of the present technology can include a step of authorizing, by the human user utilizing an interface of the base-level AI agent, the LLM to download data and automatically add the data to the training data set or to create a new training data set.

[00453] In some embodiments, the downloaded data can be provided to the human user by way of the computer system for approval by the human user prior to adding to the training data set or to the creation of the new training data set.

5 [00454] Some embodiments of the present technology can include a step of cloning the user PSI into one or more cloned user PSIs.

[00455] Some embodiments of the present technology can include a step of training each of the cloned PSIs using training data sets different from each other, wherein the training data sets of each of cloned PSIs each having weights different from each other.

10 [00456] Some embodiments of the present technology can include a step of combining the weighted training data sets from the user PSI and one or more of the cloned PSIs to create a combined training data set.

[00457] Some embodiments of the present technology can include a step of offering the any one of or any combination of the weighted training data sets of any one of or any combination of the cloned PSIs to any one of the additional AI agents and the additional PSIs for use in the training.

15 [00458] Some embodiments of the present technology can include a step of training the user PSI with the weighted training data sets of any one of or any combination of the cloned PSIs.

[00459] Some embodiments of the present technology can include a step of executing, by the human user utilizing the computer system, an automatic process of the user PSI to perform tasks without intervention by the human user unless predetermined parameters triggered.

20 [00460] Some embodiments of the present technology can include a step of monitoring, by the base-level AI agent or the user PSI, an activity against a list of prohibited activities and triggering an intervention activity that is provided to the human user.

25 [00461] Some embodiments of the present technology can include a step of detecting, by the base-level AI agent or the user PSI, if a gap in the training data sets exists, and if so, then generate interactions with the intelligent entities to obtain data that fills the gaps.

[00462] In some embodiments, the interactions between more than one of the intelligent entities can be performed in parallel.

30 [00463] Some embodiments of the present technology can include a step of providing a task to a plurality of the additional AI agents or the additional PSIs on the network, providing a result to the task from each of the additional AI agents or the additional PSIs, and determining a winning result from the results.

[00464] In some embodiments, the additional AI agents or the additional PSIs can work on the task independently and in parallel with each other.

[00465] In some embodiments, one of one or more of the additional AI agents or the additional PSIs can be un-customized, and one or more of the additional AI agents or the additional PSIs can be customized.

5 [00466] Some embodiments of the present technology can include a step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.

[00467] Some embodiments of the present technology can include a step of utilizing a universal problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.

10 [00468] Some embodiments of the present technology can include a step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.

[00469] In some embodiments, each of the additional AI agents or the additional PSIs can have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.

15 [00470] In some embodiments, records of activity can be available to any computer device on the network.

[00471] Some embodiments of the present technology can include a step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.

20 [00472] Some embodiments of the present technology can include a step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.

[00473] According to yet another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

creating a user PSI by:

acquiring a base-level AI agent that has previously been customized;

collecting media information related to the human user;

30 analyzing the media information;

transforming the analyzed media information into training data sets;

differentially weighting the transformed training data sets; and

applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;

communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics; recording all actions by the user PSI and the additional AI agents or the additional PSIs in an
5 auditable form on any one of or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs; and monitoring that each of the actions follow the set of rules, and flagging any of the actions that do not follow the set of rules.

[00474] In some embodiments, the recording of the actions can utilize blockchain technology.

10 [00475] In some embodiments, the recording of the actions can utilize a universal problem solving framework to record any one of or any combination of goals, sub-goals, problem states, and steps taken in a problem solving or cognitive activity that takes place on the network.

[00476] In some embodiments, the recorded actions can be changed only when a majority of the additional AI agents or the additional PSIs on the network provide an authorization for the change.

15 [00477] Some embodiments of the present technology can include a step of halting any of the flagged actions and applying a preventive operation to prevent a re-occurrence of the flagged action.

[00478] Some embodiments of the present technology can include a step of identifying one or more of the additional AI agents or the additional PSIs that provided the flagged action, and controlling a participation of the identified AI agent or PSI on the network.

20 [00479] Some embodiments of the present technology can include a step of analyzing the flagged action and adjusting any one of or any combination of the training data sets, the set of rules, and attributes of the network based on the analyses of the flagged action.

[00480] According to still yet another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities
25 including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

creating a user PSI by:

acquiring a base-level AI agent that has previously been customized;

collecting media information related to the human user;

30 analyzing the media information;

transforming the analyzed media information into training data sets;

differentially weighting the transformed training data sets; and

applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;

providing a task to the user PSI and to a plurality of additional AI agents or additional PSIs on the network;

providing a result to the task from each of the user PSI and the additional AI agents or the additional PSIs; and

5 determining a winning result from the results.

[00481] In some embodiments, each of the additional AI agents or the additional PSIs can work on the task independently and in parallel with each other.

[00482] In some embodiments, one of one or more of the additional AI agents or the additional PSIs can be un-customized, and one or more of the additional AI agents or the additional PSIs are
10 customized.

[00483] Some embodiments of the present technology can include a step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.

[00484] Some embodiments of the present technology can include a step of utilizing a universal
15 problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.

[00485] Some embodiments of the present technology can include a step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.

20 [00486] In some embodiments, each of the additional AI agents or the additional PSIs can have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.

[00487] In some embodiments, records of activity can be available to any computer device on the network.

25 [00488] Some embodiments of the present technology can include a step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.

[00489] Some embodiments of the present technology can include a step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.

30 [00490] In some embodiments, the network can include the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.

[00491] According to another aspect, the present technology can include a method for PSI using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities

including a human user utilizing a user computer system, additional AI agents, and additional PSIs, all electronically communicating over a collective network. The method can include the steps of:

- a) creating a user PSI by:
 - acquiring a base-level AI agent that has previously been customized;
 - 5 collecting media information related to the human user;
 - analyzing the media information;
 - transforming the analyzed media information into training data sets;
 - differentially weighting the transformed training data sets; and
 - applying the weighted transformed training data sets to the base-level AI agent to create a user
- 10 PSI, wherein the base-level AI agent;
- b) communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics;
- c) performing a baseline performance of the user PSI and the additional AI agents or the additional
- 15 PSIs utilizing a variety of standardized tasks;
- d) creating one or more versions of the user PSI each different from each other;
- e) determining a performance of the network of the user PSI, the versions of the user PSI, and the additional AI agents or the additional PSIs;
- f) assigning a credit value or a blame value to one or more elements that resulted in superior or
- 20 inferior performance of the baseline performance; and
- g) determining which of the elements are superior elements that resulted in an increase in performance and retaining the superior elements for future use.

[00492] Some embodiments of the present technology can include a step of determining which of the elements are inferior elements that resulted in a decrease in performance and change an attribute

25 of the inferior elements.

[00493] Some embodiments of the present technology can include a step of repeating step c) utilizing the superior elements and the changed inferior elements.

[00494] Some embodiments of the present technology can include a step of repeating steps c)-g) until no more superior elements are detected.

30 [00495] In some embodiments, the network can include the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.

[00496] Some embodiments of the present technology can include a step of recording all actions by the user PSI and the additional AI agents or the additional PSIs in an auditable form on any one of

or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs.

[00497] In some embodiments, the standardized tasks can be any one of or any combination of problem solving tasks using a shared universal problem solving architecture, ethical and safety scenarios designed to determine whether any one of or any combination of the use PSI, and the additional AI agents or the additional PSIs behaves safely and ethically, standardized intelligence tests and assessments that have been designed to test an intelligence of the intelligent entities, tasks configured to measure a degree of “hallucination” or production of erroneous results, tasks that come from a variety of domains of expertise, tasks that incorporate different cultural or group norms for behavior, tasks that are random permutations of existing standardized tasks, and tasks that are dynamically created based on changing conditions.

[00498] In some embodiments, the versions of the user PSI can be created by any one of or any combination of: human adjustment of parameters, weights or data encoding a knowledge of the user PSI; autonomous adjustment of the parameters, weights or data encoding the knowledge the user PSI; random variation of the parameters, weights or data encoding the knowledge of the user PSI; subjecting the user PSI to different training regimes or amounts of training; creating the different versions of the user PSI sequentially, and creating the different versions of the user PSI in parallel.

[00499] In some embodiments, the versions of the user PSI can be created by directly combining the parameters, weights or data encoding the knowledge of multiple PSIs by computing average weight values, with weighting more recently created or more complex versions of the user PSI more than older or less complex versions of the user PSI.

[00500] In some embodiments, the versions of the user PSI can be created by random or deliberate methods for estimating which of the versions are to have beneficial results, wherein the methods for estimating being any one of or any combination of comparing a degree of match between a knowledge of one of the versions of the user PSI and a statistical frequency of tasks that are submitted to the network, comparing an overlap in knowledge of the additional PSIs that are different on the network such that changes are made to optimize a performance of groups of the versions of the user PSI given a characterization of an entire network knowledge compared to a characteristics of a problem that the network is expected to solve, and using hill climbing or gradient descent for optimizing parameters of any one of the versions of the user PSI.

[00501] According to another aspect, the present technology can include a method for PSI utilizing a single computerized intelligent system including multiple AI agents residing in the single computerized intelligent system. The method can include:

acquiring a base-level AI agent that has previously been customized, the base-level AI agent resides in a single computerized intelligent system;
collecting media information related to a human user associated with the base-level AI agent;
transforming the analyzed media information into training data sets;
5 differentially weighting the transformed training data sets;
applying the weighted transformed training data sets to the base-level AI agent to create a user PSI; and
communicating the user PSI with multiple additional PSIs to enable community-based safety features from the additional PSIs to the user PSI.
10 training a base Large Language Model (LLM) of an AI agent with guardrails including attributes associated with any one of or any combination safety, ethics and knowledge, the AI agent residing in a single computerized intelligent system;
customizing the base LLM to an ethics profile;
combining ethical information from multiple additional AI agents residing in the single
15 computerized intelligent system, the additional AI agents being different to that of the AI agent, the additional AI agents residing in the single computerized intelligent system;
refining a set of values of the base LLM based on problem solving of a problem request; and
updating the base LLM with the combined ethical information and the refined set of values thereby allowing for a scalable AGI.

20

Concluding Remarks

[00502] The present technology has attempted to explain some of the ways that we can efficiently and effectively design next steps in the development of SuperIntelligence, PSIs, and “Community SuperIntelligence”. Humans have the ability NOW to make design decisions that greatly affect the
25 future trajectory of Planetary Intelligence. We must design PSIs, networks of PSIs, and other AI systems with humans in the loop (initially). These systems must serve human values (even when the intelligence outstrips that of humans). A Community of PSIs approach can help ensure stable, human-centered values even when the pace of growth of PSIs vastly outstrips the ability of humans to keep up intellectually. If we design such systems correctly, based on the principles outlined in this patent now, the future can be an amazingly wonderful place for all humanity, and all sentient beings.

30 **[00503]** While embodiments of the safe PSI have been described in detail, it should be apparent that modifications and variations thereto are possible, all of which fall within the true spirit and scope of the present technology. With respect to the above description then, it is to be realized that the optimum relationships for the parts of the present technology, to include variations in size,

materials, shape, form, function and manner of operation, assembly and use, are deemed readily apparent and obvious to one skilled in the art, and all equivalent relationships to those illustrated in the drawings and described in the specification are intended to be encompassed by the present technology. For example, any suitable implementation may be used instead of the above-described.

- 5 [00504] Therefore, the foregoing is considered as illustrative only of the principles of the present technology. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the present technology to the exact construction and operation shown and described, and accordingly, all suitable modifications and equivalents may be resorted to, falling within the scope of the present technology.

CLAIMS

What is claimed is:

1. A system for personalized super intelligence (PSI) using intelligent agents to develop and continuously improve PSI for a human user utilizing a computer system, and additional PSIs, all electronically communicating over a collective network, the system comprising:
 - a computer system comprising: a processor, a computer-readable storage medium, and program instructions stored on the computer-readable storage medium being executable by the processor to cause the computer system to:
 - implement a base-level Artificial Intelligence (AI) agent on the computer system, wherein
 - the base-level AI agent has already been customized;
 - collect media information related to the human user;
 - analyze the media information;
 - transform the analyzed media information into training data sets;
 - differentially weighting the transformed training data sets;
 - add knowledge modules to the weighted transformed training data sets;
 - locate new sources of data to include to the weighted transformed training data sets;
 - apply the weighted transformed training data sets to the base-level AI agent to create a personalized PSI; and
 - communicate the personalized PSI with multiple additional PSIs using the network to enable community-based safety features from multiple additional PSIs to the personalized PSI.
2. A method for personalized super intelligence (PSI) using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a computer system, additional Artificial Intelligence (AI) agents, and additional PSIs, all electronically communicating over a collective network, the method comprising the steps of:
 - acquiring a base-level AI agent that has previously been customized;
 - collecting media information related to the human user;
 - analyzing the media information;
 - transforming the analyzed media information into training data sets;
 - differentially weighting the transformed training data sets;
 - adding knowledge modules to the weighted transformed training data sets;
 - locating new sources of data to include to the weighted transformed training data sets;
 - applying the weighted transformed training data sets to the base-level AI agent to create a user PSI; and

communicating the user PSI with multiple additional PSIs using the network to enable community-based safety features from the additional PSIs to the user PSI.

3. The method of claim 2 further comprising the step of obtaining, by the human user, by the base-level AI agent or by the user PSI, any one of or any combination of new training datasets and training modules to be added to the weighted transformed training data sets.

4. The method of claim 2 further comprising the step of monetizing the user PSI by enabling other human users or other PSIs to access and use the weighted transformed training data sets of the personalized PSI.

5. The method of claim 2, wherein the base-level AI agent is any one of a pre-trained Large Language Model (LLM), other tunable and trainable AI agent, and one or more customized AIs from other human owners.

6. The method of claim 2, wherein the media information is any one of or any combination of videos of the human user, videos of people and topics related to the human user, photographs or images of the human user, photographs or images of people and topics related to the human user, writings related to the human user, journals related to the human user, blogs related to the human user, posts related to the human user, tweets related to the human user, emails related to the human user, podcasts related to the human user, recordings related to the human user, auditory content related to the human user, auditory content of people and topics related to the human user, data collected by third-party vendors that are related to the human user, websites related to the human user, apps related to the human user, online information related to the human user, social media information related to the human user, and other AI agents related to the human user.

7. The method of claim 2 further comprising the step of providing permission to one or more social media platforms allowing social media content related to the human user to be accessible by the base-level AI agent or the user PSI.

8. The method of claim 2, wherein the analyzing of the media information further includes the annotating and categorizing the media information.

9. The method of claim 8, wherein the analyzing of the media information utilizes one or more algorithms including any one of or any combination of transformer algorithms, deep learning algorithms, transcription algorithms, content and sentiment analysis and summarization, LLMs, and crowdsourced and crowd-supervised human.

10. The method of claim 2 further comprising the step of compensating, by the human user, one or more human workers on a crowd-sourcing website to review the training data sets, or to suggest refinements to the training data sets.

11. The method of claim 2, wherein the step of weighting the transformed training data sets is performed by adjusting a weight value of any one of the transformed training data sets by the human user utilizing an interface on the computer system that is implementing the base-level AI agent.
12. The method of claim 2, wherein the step of weighting the transformed training data sets is performed by any one of or any combination of the additional AI agents, the additional PSIs, and additional human workers each utilizing a computer system.
13. The method of claim 2 further comprising the step of inputting ethical values to the training data sets by providing a series of interactive dialogs to the human user, the interactive dialogues include predetermined ethical scenarios.
14. The method of claim 2 further comprising the step of mixing the weights of the transformed training data sets to improve the LLM of the base-level AI agent.
15. The method of claim 2 further comprising the step of authorizing, by the human user utilizing an interface of the base-level AI agent, the LLM to download data and automatically add the data to the training data set or to create a new training data set.
16. The method of claim 15, wherein the downloaded data is provided to the human user by way of the computer system for approval by the human user prior to adding to the training data set or to the creation of the new training data set.
17. The method of claim 2 further comprising the step of cloning the user PSI into one or more cloned user PSIs.
18. The method of claim 17 further comprising the step of training each of the cloned PSIs using training data sets different from each other, wherein the training data sets of each of cloned PSIs each having weights different from each other.
19. The method of claim 18 further comprising the step of combining the weighted training data sets from the user PSI and one or more of the cloned PSIs to create a combined training data set.
20. The method of claim 18 further comprising the step of offering the any one of or any combination of the weighted training data sets of any one of or any combination of the cloned PSIs to any one of the additional AI agents and the additional PSIs for use in the training.
21. The method of claim 18 further comprising the step of training the user PSI with the weighted training data sets of any one of or any combination of the cloned PSIs.
22. The method of claim 2 further comprising the step of executing, by the human user utilizing the computer system, an automatic process of the user PSI to perform tasks without intervention by the human user unless predetermined parameters are triggered.

23. The method of claim 2 further comprising the step of monitoring, by the base-level AI agent or the user PSI, an activity against a list of prohibited activities and triggering an intervention activity that is provided to the human user.
24. The method of claim 2 further comprising the step of detecting, by the base-level AI agent or the user PSI, if a gap in the training data sets exists, and if so, then generating interactions with the intelligent entities to obtain data that fills the gaps.
25. The method of claim 24, wherein the interactions between more than one of the intelligent entities is performed in parallel.
26. The method of claim 2 further comprising the step of providing a task to a plurality of the additional AI agents or the additional PSIs on the network, providing a result to the task from each of the additional AI agents or the additional PSIs, and determining a winning result from the results.
27. The method of claim 26, wherein each of the additional AI agents or the additional PSIs works on the task independently and in parallel with each other.
28. The method of claim 26, wherein one of one or more of the additional AI agents or the additional PSIs are un-customized, and one or more of the additional AI agents or the additional PSIs are customized.
29. The method of claim 26 further comprising the step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.
30. The method of claim 26 further comprising the step of utilizing a universal problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.
31. The method of claim 2 further comprising the step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.
32. The method of claim 2, wherein each of the additional AI agents or the additional PSIs have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.
33. The method of claim 2, wherein records of activity is available to any computer device on the network.
34. The method of claim 2 further comprising the step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.
35. The method of claim 2 further comprising the step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.

36. A method for personalized super intelligence (PSI) using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional Artificial Intelligence (AI) agents, and additional PSIs, all electronically communicating over a collective network, the method comprising the steps of:

- 5 creating a user PSI by:
 - acquiring a base-level AI agent that has previously been customized;
 - collecting media information related to the human user;
 - analyzing the media information;
 - transforming the analyzed media information into training data sets;
 - 10 differentially weighting the transformed training data sets; and
 - applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;
- 15 communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics;
- 20 recording all actions by the user PSI and the additional AI agents or the additional PSIs in an auditable form on any one of or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs; and
- monitoring that each of the actions follow the set of rules, and flagging any of the actions that do not follow the set of rules.
37. The method of claim 36, wherein the recording of the actions utilizes blockchain technology.
38. The method of claim 36, wherein the recording of the actions utilizes a universal problem solving framework to record any one of or any combination of goals, sub-goals, problem states, and steps taken in a problem solving or cognitive activity that takes place on the network.
- 25 39. The method of claim 36, wherein the recorded actions are changed only when a majority of the additional AI agents or the additional PSIs on the network provide an authorization for the change.
40. The method of claim 36 further comprising the step of halting any of the flagged actions and applying a preventive operation to prevent a re-occurrence of the flagged action.
41. The method of claim 40 further comprising the step of identifying one or more of the additional
- 30 AI agents or the additional PSIs that provided the flagged action and controlling a participation of the identified AI agent or PSI on the network.
42. The method of claim 36 further comprising the step of analyzing the flagged action and adjusting any one of or any combination of the training data sets, the set of rules, and attributes of the network based on the analyses of the flagged action.

43. A method for personalized super intelligence (PSI) using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional Artificial Intelligence (AI) agents, and additional PSIs, all electronically communicating over a collective network, the method comprising the steps of:

- 5 creating a user PSI by:
- acquiring a base-level AI agent that has previously been customized;
 - collecting media information related to the human user;
 - analyzing the media information;
 - transforming the analyzed media information into training data sets;
 - 10 differentially weighting the transformed training data sets; and
 - applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;
 - providing a task to the user PSI and to a plurality of additional AI agents or additional PSIs on the network;
 - 15 providing a result to the task from each of the user PSI and the additional AI agents or the additional PSIs; and
 - determining a winning result from the results.

44. The method of claim 43, wherein each of the additional AI agents or the additional PSIs works on the task independently and in parallel with each other.

- 20 45. The method of claim 43, wherein one of one or more of the additional AI agents or the additional PSIs are un-customized, and one or more of the additional AI agents or the additional PSIs are customized.

46. The method of claim 43 further comprising the step of utilizing training data sets of the additional AI agents or the additional PSIs with the winning result in the customization of the user base-level AI agent or the user PSI.

- 25 47. The method of claim 43 further comprising the step of utilizing a universal problem solving architecture by each of the additional AI agents or the additional PSIs on the task, respectively.

48. The method of claim 43 further comprising the step of using a set of safety or ethical rules that each of the user base-level AI agent, the user PSI, the additional AI agents and the additional PSIs agrees to use.

- 30 49. The method of claim 43, wherein each of the additional AI agents or the additional PSIs have a set of weights that encode a knowledge of the additional AI agents and the additional PSIs, respectively.

50. The method of claim 43, wherein records of activity is available to any computer device on the network.
51. The method of claim 43 further comprising the step of providing a task to the user PSI that is configured to create a hallucination by an LLM of the PSI.
- 5 52. The method of claim 43 further comprising the step of providing a task to the user PSI, wherein the task is a random permutation of an existing standardized task, or a dynamically created task based on changing conditions.
53. The method of claim 43, wherein the network includes the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.
- 10 54. A method for personalized super intelligence (PSI) using intelligent entities to develop and continuously improve PSI, wherein the intelligent entities including a human user utilizing a user computer system, additional Artificial Intelligence (AI) agents, and additional PSIs, all electronically communicating over a collective network, the method comprising the steps of:
- a) creating a user PSI by:
 - 15 acquiring a base-level AI agent that has previously been customized;
 - collecting media information related to the human user;
 - analyzing the media information;
 - transforming the analyzed media information into training data sets;
 - differentially weighting the transformed training data sets; and
 - 20 applying the weighted transformed training data sets to the base-level AI agent to create a user PSI, wherein the base-level AI agent;
 - b) communicating the user PSI with multiple additional AI agents or additional PSIs using the network to enable community-based safety features, wherein the user PSI and the additional AI agents or the additional PSIs each agree to use a set of rules relating to safety or ethics;
 - 25 c) performing a baseline performance of the user PSI and the additional AI agents or the additional PSIs utilizing a variety of standardized tasks;
 - d) creating one or more versions of the user PSI each different from each other;
 - e) determining a performance of the network of the user PSI, the versions of the user PSI, and the additional AI agents or the additional PSIs;
 - 30 f) assigning a credit value or a blame value to one or more elements that resulted in superior or inferior performance of the baseline performance; and
 - g) determining which of the elements are superior elements that resulted in an increase in performance and retaining the superior elements for future use.

55. The method of claim 54 further comprising the step of determining which of the elements are inferior elements that resulted in a decrease in performance and changing an attribute of the inferior elements.
56. The method of claim 55 further comprising the step of repeating step c) utilizing the superior
5 elements and the changed inferior elements.
57. The method of claim 55 further comprising the step of repeating steps c)-g) until no more superior elements are detected.
58. The method of claim 54, wherein the network includes the user PSI and the additional AI agents or the additional PSIs, and one or more additional networks each including AI agents or PSIs.
- 10 59. The method of claim 54 further comprising the step of recording all actions by the user PSI and the additional AI agents or the additional PSIs in an auditable form on any one of or any combination of a central computer system on the network, the user computer system, the additional AI agents and the additional PSIs.
60. The method of claim 54, wherein the standardized tasks are any one of or any combination of
15 problem solving tasks using a shared universal problem solving architecture, ethical and safety scenarios designed to determine whether any one of or any combination of the use PSI, and the additional AI agents or the additional PSIs behaves safely and ethically, standardized intelligence tests and assessments that have been designed to test an intelligence of the intelligent entities, tasks configured to measure a degree of "hallucination" or production of erroneous results, tasks that
20 come from a variety of domains of expertise, tasks that incorporate different cultural or group norms for behavior, tasks that are random permutations of existing standardized tasks, and tasks that are dynamically created based on changing conditions.
61. The method of claim 54, wherein the versions of the user PSI are created by any one of or any combination of: human adjustment of parameters, weights or data encoding a knowledge of the user
25 PSI; autonomous adjustment of the parameters, weights or data encoding the knowledge the user PSI; random variation of the parameters, weights or data encoding the knowledge of the user PSI; subjecting the user PSI to different training regimes or amounts of training; creating the different versions of the user PSI sequentially, and creating the different versions of the user PSI in parallel.
62. The method of claim 54, wherein the versions of the user PSI are created by directly combining
30 the parameters, weights or data encoding the knowledge of multiple PSIs by computing average weight values, with weighting more recently created or more complex versions of the user PSI more than older or less complex versions of the user PSI.
63. The method of claim 54, wherein the versions of the user PSI are created by random or deliberate methods for estimating which of the versions are to have beneficial results, wherein the

methods for estimating being any one of or any combination of comparing a degree of match between a knowledge of one of the versions of the user PSI and a statistical frequency of tasks that are submitted to the network, comparing an overlap in knowledge of the additional PSIs that are different on the network such that changes are made to optimize a performance of groups of the versions of the user PSI given a characterization of an entire network knowledge compared to a characteristics of a problem that the network is expected to solve, and using hill climbing or gradient descent for optimizing parameters of any one of the versions of the user PSI.

64. A method for personalized super intelligence (PSI) utilizing a single computerized intelligent system including multiple Artificial Intelligence (AI) agents residing in the single computerized intelligent system, the method comprising:

acquiring a base-level AI agent that has previously been customized, the base-level AI agent resides in a single computerized intelligent system;
collecting media information related to a human user associated with the base-level AI agent;
transforming the analyzed media information into training data sets;
differentially weighting the transformed training data sets;
applying the weighted transformed training data sets to the base-level AI agent to create a user PSI; and
communicating the user PSI with multiple additional PSIs to enable community-based safety features from the additional PSIs to the user PSI.
training a base Large Language Model (LLM) of an AI agent with guardrails including attributes associated with any one of or any combination safety, ethics and knowledge, the AI agent residing in a single computerized intelligent system;
customizing the base LLM to an ethics profile;
combining ethical information from multiple additional AI agents residing in the single computerized intelligent system, the additional AI agents being different to that of the AI agent, the additional AI agents residing in the single computerized intelligent system;
refining a set of values of the base LLM based on problem solving of a problem request; and
updating the base LLM with the combined ethical information and the refined set of values thereby allowing for a scalable AGI.

30