SUPERINTELLIGENCE
AN iQ COMPANY

SUPERINTELLIGENCE DESIGN WHITE PAPER #3: SYSTEM AND METHODS FOR HUMAN-CENTERED AGI

by Dr. Craig A. Kaplan
May 2025

Note: To provide as much information on our designs and inventions for safe AGI and SuperIntelligence as quickly as possible, the following white paper text currently consists of the descriptions of inventions and designs that have not yet been formatted according to conventional standards for journal publication. As time allows, these descriptions will be revised and updated to include more traditional formatting, including additional references. All diagrams will be made available in a separate file. Meanwhile, we hope that the description in this white paper will help researchers and developers pursue safer, faster, and more profitable approaches to developing advanced AI, AGI, and SI systems that reduce p(doom) for all humanity.

SUPERINTELLIGENCE
AN iQCOMPANY

TABLE OF CONTENTS

ABSTRACT

The safest path to AGI must also be the fastest for it to be effective.

The preferred implementation of AGI is the fastest method for achieving AGI because it begins with a network of human problem-solving agents, who, by definition, can perform any intellectual task as well or better than the average human. AI agents that have been trained and customized by individual humans (AAAIs) are introduced to the network as AI problem-solving agents. The human and AI agents share a common problem-solving architecture: rigorous, scalable, transparent, auditable, safe, and powerful.

The architecture supports automatic learning and self-improvement. It is compatible with LLMs, which can be “plugged in” to the network and upgraded as more powerful LLM models become available. The AGI network begins with humans doing most of the problem-solving work, especially the most important aspects, such as setting goals, and the most difficult aspects, such as representing the problem. Over time, AAAIs do more and more of the actual work, more effectively and efficiently than humans could, while human attention is increasingly directed to issues of ethics, safety, and oversight.

Because ethics and safety checks are built into the architecture itself, as the speed of problems increases far beyond the capability of humans to “keep pace,” the system remains aligned with human values and ethics. At any time, humans can see precisely how the system makes decisions, including all ethical information.

The AGI network is highly scalable and will become more powerful over time, yet the fundamental values and ethics of the system, which cannot be logically derived by any intelligence, no matter how smart and fast, remain aligned with humans. Thus, the invention solves the alignment problem in a democratic and scalable manner.

The fact that the AGI network can be implemented rapidly – far faster than estimates for when AGI will develop from other approaches – ensures a first mover advantage that allows this safest path to AGI to dominate other approaches thereby fulfilling the two key requirements for the invention, namely that it be not only the fastest path to AGI but also the safest.

BACKGROUND

This Provisional Patent Application incorporates by reference all the material in earlier Provisional Patent Applications #63/487,494 and 63/491/040.

AGI is AI that can perform any intellectual task as well or better than the average human being. Eric Schmidt, just about a year before the date of this PPA, said the median estimate for when AGI would be developed that is also capable of setting its own goals or objectives was 2042. Ray Kurzweil, a well-known futurist, puts the date earlier in 2029. With the rapid adoption of GPT, the ensuing AI arms race, and the movement of all major companies towards generative AI, the date for AGI has advanced. The commercial and geopolitical pressure to be first has increased. The invention described in this PPA would enable AGI to become operational by 2025, much faster than most predict. How is this possible when most researchers have no clear idea of achieving AGI?

Simply put, most of the AI research community is looking at the problem incorrectly.

The typical research approach is to engineer ever-more-powerful Large Language Models (LLMs) at great expense until one exceeds human intelligence and intellectual capabilities in all areas. Then, we have an “alignment problem” where the humans worry about whether the goals of this SuperIntelligent AI “align” with human values.

Misalignment could mean the extinction of the human species. We hope that won’t happen. We used to think we had until 2042 to figure out how to make things safe.

Some thought a few of these SuperIntelligent AGIs would be owned by large countries and protected in the way Plutonium is protected. The reasoning, logical if one accepts the standard model of building ever more powerful LLMs, was that only a few countries would have the vast computational resources needed to build an AGI. Thus, like nuclear material, it could remain accessible to most and a carefully guarded secret. With luck, all the powerful countries could somehow guard and contain the SuperIntelligent AGIs, and the human species would be preserved.

Unfortunately, no. That thinking is erroneous. Source code for powerful LLMs is already open-sourced and adopted by hundreds of millions of people. Anyone can trick GPT or any LLM into saying or potentially doing bad things (since these LLMs are now connected to the web and capable of programming). AgentGPT and other systems (e.g., those using Langchain and/or other code that extends the abilities of LLMs and allows them to set their own goals and subgoals using software techniques well known in the art) are already setting and pursuing them.

Few truly understand how close humanity is to extinction at this juncture. A survey conducted BEFORE the release of ChatGPT had 48% putting the risk of human extinction by AI at 10% or greater. In May 2023, I estimate the chances of extinction at 20%. That's like playing Russian Roulette with 8 billion lives and a five-shot revolver.

Humankind deserves better than this, and fortunately, a far superior answer exists, as described in this patent. First, let's state some facts that almost all the top AI researchers and heads of leading AI companies seem to agree on:

1. At SOME POINT, AGI will develop and be capable of setting its own goals.
2. The "alignment problem" is a genuine concern.
3. The power involved in AGI means that regulating it will not protect humans. Regulation will only slow down some countries or companies, enabling others to gain an advantage.

Faced with these facts, the message to the general public adopted by most AI thought leaders has been: "No need to panic. The dawn of SuperIntelligent AGI is far away. We will figure something out." Inwardly, they are terrified or in denial. In Max Tegmark's words, none of that stops them from racing forward, not an AI arms race but a "suicide race."

Since no one knows how to create SuperIntelligent AGI, and no one understands what makes the current LLMs behave and "reason" as they do, each company is investing ever larger sums of money into trying to be the first with AGI. Simultaneously, each spends time and resources worrying about ensuring it doesn't kill us all. And everyone gives lip service to "Responsible AI." This is the situation as of May 2023, as near as I can surmise. Humanity does indeed deserve better.

THE FALSE ASSUMPTION

The false assumption made by almost all AI researchers is that AGI will develop as the result of training an Uber-LLM. Humans may be involved in the initial training and supervision, but after that, it will train itself, write its own code, and ultimately set its own goals. At this point, we have a potential "alignment problem."

WHAT IF… AGI could be created now? What if the AGI that was created now had humans in the loop and gave humans the democratic opportunity to transfer their values to the AGI? Then we would have a situation where: 1) AGI was created years earlier than anyone envisions, and 2) there would be no alignment problem. AGI would be safer than the "wait and hope" approaches that training an Uber-LLM leaves us with.

THE FASTEST PATH MUST BE THE SAFEST PATH

Maybe the alignment problem should be called the “end of the humanity problem.” Whichever AGI is achieved first is the one we have to worry about. Since SuperIntelligent AGI (“AGI” for brevity) can improve itself at an exponential rate, the first AGI could theoretically dominate all the others if it had a sufficient head start and if the other runner-up AGIs did not have superior (e.g., faster) learning algorithms enabling them to pull ahead. This is potentially a winner-take-all scenario.

WINNER-TAKE-ALL

Few things in life or business are winner-take-all. The idea of “first mover advantage” is common to Silicon Valley venture capitalists, but they know that being first doesn’t make you best. Facebook beat Friendster, and Friendster was first. Xerox Parc was first with the Graphical User Interface, but Steve Jobs took it, and Apple is now the world’s largest company, while Xerox is all but forgotten. Usually, being first is an advantage, but being bigger is better. Most of the time, you don’t need to be first to win. And the winner rarely takes all. iPhones have competition.

But this time is different. You should be skeptical of those words.

But that skepticism should not close your mind entirely to a logical and well-reasoned argument. Simply put, Artificial Intelligence will be more intelligent than us, able to set its own goals, (re)program itself, and learn exponentially by creating billions of copies of itself and having these copies improve each other.

AI starts as a tool, yes, but it will not remain one. It will become an intelligent entity trillions of times smarter, faster, and more perceptive than us. We have never created such an entity before. Whichever version gets a head start, assuming it maintains the fastest learning rate, will dominate all other AGIs and incorporate them into its intelligence. That is the most likely outcome. So, AGI is likely winner-take-all.

In such a scenario, we have to worry about two things: 1) How to create AGI first, and 2) How to create a safe AGI. Without both conditions being met, humanity is at risk.

ENVISION AGI NOW

Part of the problem AI researchers face is that they can't envision what AGI would be, other than an Uber LLM that is years away. Instead of looking to the future, we must realize that AGI can exist NOW. We need to imagine in detail what that AGI, which can exist now, looks like and how it functions.

There may be multiple instantiations of AGI, but this patent describes the preferred implementation, which has as its chief benefits:

1. It is the fastest path to AGI, and
2. It is the safest path to AGI.

HUMAN-CENTERED AGI

From a safety perspective, many researchers would probably agree that having "humans in the loop" in any advanced technology system is good. Indeed, in the case of nuclear weapons – arguably the most powerful technology developed by humans to date – humans in the loop have saved the planet from nuclear annihilation.

For example, at least once, a Russian Colonel overrode the faulty Russian computers that said the USA had launched a nuclear attack. Had the Colonel blindly followed protocol, as a machine would have done, humanity would have experienced a nuclear holocaust. Fortunately, his human values and reasoning led him to believe the computer system was at fault (which it was), and he defied standing orders to protect millions of innocent lives. He should be a hero, and probably would have been, except that to honor him, he would have spotlighted the fact that the Russian systems were defective, and so instead, he got a pension and was quietly retired. Here's Wikipedia on it:

Stanislav Yevgrafovich Petrov (Russian: Станисла́в Евгра́фович Петро́в; 7 September 1939 – 19 May 2017) was a lieutenant colonel of the Soviet Air Defence Forces who played a key role in the 1983 Soviet nuclear false alarm incident.[1] On 26 September 1983, three weeks after the Soviet military had shot down Korean Air Lines Flight 007, Petrov was the duty officer at the command center for the Oko nuclear early-warning system when the system reported that a missile had been launched from the United States, followed by up to five more. Petrov judged the reports to be a false alarm.[2]

His subsequent decision to disobey orders, against Soviet military protocol [3], is credited with having prevented an erroneous retaliatory nuclear attack on the United States and its NATO allies that could have resulted in a large-scale nuclear war, which could have wiped

out half of the population of the countries involved. An investigation later confirmed that the Soviet satellite warning system had indeed malfunctioned. Because of his decision not to launch a retaliatory nuclear strike amid this incident, Petrov is often credited with " saving the world".[4][5][6]

“AGI is more dangerous than nuclear weapons… by a lot,” Elon Musk has said publicly. He’s right. Much more dangerous. We need humans in the loop and human values to be adopted by AGI. But most AI researchers can’t see how this will be done. Because they envision AGI arising from an Uber-LLM that trains itself and then (re)programs itself millions or trillions of times faster than the human mind can comprehend, they can’t see how humans can be “in the loop.” This situation worries them, but they can’t see a way out. It seems inevitable. It isn’t.

AI researchers have it backward. We shouldn’t be thinking about how to build an Uber AGI system and then tack on human values and safety concerns after the fact in a desperate hope that we avoid the alignment problem. Instead, we should be designing a human collective intelligence system that incorporates AI, little by little, and trains the AI on human values.

Over time, the human-AI system becomes SuperIntelligent and is powered more by AI and less by human thinking. But since humans developed it, and the AI learned values from humans, and humans (even at the AGI stage) are still part of the system, humans never left the loop, and human values and ethics are built into the DNA of the AGI system. More importantly, this system can be built immediately. It can be FIRST to win the AGI “arms race.” The solution: AGI, which we can develop first, and is also safe. Alignment problem solved!

HOW TO BUILD AGI IN THE FASTEST, SAFEST MANNER

The inventor has spent his entire professional career building intelligent systems focusing on human collective intelligence systems. At the last company he founded, he proved that it is possible to harness the brainpower of millions of average humans (retail investors) to perform as well or better than the best hedge funds on Wall Street. Two heads are better than one. It turns out two million heads, properly harnessed – even in a straightforward system – are better than the best and most highly paid minds on the planet who devote their every waking moment to trying to get an edge in the markets.

Ask yourself: How do I build a “network” that can solve any problem or do any intellectual activity better than the average human?

One answer is to build a network of human problem solvers, combined with a universal problem-solving architecture that allows them to work together in a coordinated and rigorously defined way. A problem is submitted to the network. One or more human problem solvers work on the problem and return a solution. Since the network includes humans, it can solve any problem the average human can solve. And since there are many humans, if the efforts are intelligently coordinated, then the network will often perform better than the average human, just as in my prior company (PredictWallStreet) millions of average humans were able to beat most Wall Street pros at the extremely difficult problem of getting an edge in the financial markets.

So far, we have established that a network of HUMANS can perform as well as the average human on any problem, and likely better. But what does that have to do with AGI? Isn't the whole point of AGI that AI scales and thinks much faster than humans, while humans, especially if we have to include the time needed for communication and coordination between humans, will think much slower? Wouldn't a network of humans scale very poorly? Isn't that why we need an Uber LLM to be the AGI? These are all reasonable questions addressed in this description of the invention.

COLLECTIVE INTELLIGENCE OF HUMAN AND AI PROBLEM SOLVERS

First, consider that the "solvers" on the network don't have to be ALL human solvers. They can be AI solvers too. LLMs exist, and it is possible to include the latest generation of LLMs as solvers participating in the network. So now we are talking about a hybrid network of human and AI problem solvers. They need a common language for problem solving, and that language has to be rigorous. It needs to be rigorous because machines are quite literal, and if the coordination between humans and AI is loosely specified, there is more opportunity for error. This is particularly true since, as Eric Schmidt has pointed out, humans make many assumptions when they work together that might not be valid for a machine.

Humans know, for instance, roughly the intellectual and ethical limitations under which other humans work. No human will propose a solution that wipes out themselves and all of humanity. That would be counterproductive. Humans might be fearful and greedy at times, but they generally are not suicidal.

SOCIOPATHIC AI

Machines have no such constraints. A machine might very well come up with “solutions” that are detrimental to humans simply because a machine is not human and either doesn’t know better or doesn’t care. Most humans (sociopaths excepted) have empathy and care to some degree about other humans.

Machines are neither good nor evil inherently. AI agents are “sociopathic” in the sense that they have intellectual abilities equal to or surpassing humans, without the corresponding levels of empathy. That is part of the reason so many leading AI thinkers are worried about the Alignment problem. They should worry and then turn that worry into positive action!

TWO REASONS FOR A RIGOROUS ARCHITECTURE

One reason we need a rigorously specified problem-solving architecture that can coordinate the actions of not only human problem solvers but also AI problem solvers (collectively “solvers”) is that the AIs won’t necessarily use “common sense” and “empathy” when participating unless the problem-solving path is explicitly delineated. Humans are involved in correcting the AI solvers when they go astray.

However, there is another, more technical reason for a rigorously specified problem-solving protocol that both human and AI solvers follow. If problem-solving is described rigorously, then it is possible to design a system that learns various solutions as they are developed.

Moreover, the learned solutions are auditable and entirely understandable for humans, something that current LLM intelligence lacks, which is very worrisome to humans. If we cannot know how the LLM arrives at a decision or solution, how can we trust it? How can we know it is “safe” for humans?

With the collective solver network approach to AGI, every solution path is rigorously specified and known, which not only avoids ethical errors but also provides auditable transparency and facilitates learning. AI is very good at learning rigorously specified things and less good at learning vague and unstructured things.

BENEFITS OF A COMMON ARCHITECTURE FOR AI AND HUMAN COGNITION

The benefits of a preferred implementation of a rigorously specified typical architecture for AI and human cognition, at least with regard to coordinated problem-solving on a network of human and AI agents, will include, without limitation:

1. Avoids unintentional error due to loose specifications
2. Enables automatic learning of rigorous solutions
3. Enables scalability to any problem or intellectual endeavor
4. Enables modularity and stability
5. Maximizes safety

We shall briefly expand on each of these desired benefits and then explain the preferred implementation of a typical Architecture of Cognition that achieves them.

AVOIDING UNINTENTIONAL ERRORS

First, we have already mentioned that because humans and AIs don't think the same way, loose specifications can lead to error. Constraints that any human would understand, such as "don't implement a solution that ends all of humanity", might seem perfectly acceptable to a machine if the humans were not "in the loop" to set the machine straight. Rigorously specifying what is, and is not, an ethical solution (for example) requires that the standard architecture for problem-solving be rigorously specified.

ENABLING AUTOMATIC LEARNING

Second, the more precisely specified a solution is, the easier it is for an AI to learn. While LLMs can learn from vast amounts of unstructured text and input via Transformers and other deep learning techniques, these techniques are costly, time-consuming, and impractical for learning specific chunks of knowledge, such as solutions to specific problems. On the other hand, rigorously specifying problem-solving behavior in a traceable and auditable way, such that an AI can review the steps in the solution, understand why each step was taken, and learn when to re-use that solution, is a much more practical and inexpensive approach to incremental, automated learning.

ENABLING SCALABILITY

Third, a truly general architecture for cognition allows for the representation of any problem or intellectual task humans do. The generality of such a representation means that it is truly scalable and applicable to any human intellectual endeavor – a key requirement for AGI.

ENABLING MODULARITY AND SCALABILITY

Fourth, a common architecture for cognition means that intelligent agents with a vast range of differing intellectual abilities can be “plugged in” to the network as long as they all speak the common language of the architecture. Humans have individual differences in intelligence, skills, expertise, values, and other intellectual attributes, yet we can work together.

An architecture that can accommodate a wide range of human solvers can also accommodate a wide range of AI solvers. In the future, ever more sophisticated LLMs will be developed. This preferred implementation of AGI does not discourage such efforts but instead embraces them. LLMs and the development of ever-more-powerful narrow AI systems, as well as general AI systems, are all entirely complementary for this inventive approach to AGI.

Just as human solvers with varying degrees of intelligence and skills can plug into the network, so too, different AIs with varying degrees of intelligence and abilities can also plug in. As long as all solvers follow the typical architecture, which coordinates every entity’s intellectual effort, modular intelligences with different capabilities only increase and enhance the power of the AGI network.

Further, since the behavior of all solvers is rigorously captured and described, the system is stable, and all the entities can learn from each other. AI can learn from humans; humans can learn from AI; AI can learn from AI. In all cases, the modularity and stability of the system are maintained, and the power of the AGI network increases.

MAXIMIZING SAFETY

Finally, a rigorous, universal architecture of cognition maximizes the safety (from a human standpoint) of the AGI network. One of the problems with current deep learning approaches to AI, with the idea of creating ever-more-powerful LLMs, is that the resulting LLMs are untrustworthy because humans are unable to know exactly why they behave as it does. In effect, the LLM / deep learning approach results in “alien” and potentially “sociopathic” intelligence, which quite rightly alarms thoughtful AI researchers. While it is possible, or even likely, that such alien intelligence is benign or even beneficial towards humans, without understanding how it thinks and how it reaches its conclusions, it is difficult to trust it.

Some researchers suggest that within the next five years or so, improved LLMs will be able to explain their reasoning to humans. Even if that is so, humans will still have to trust the explanations of the LLMs. How can humans be sure the LLMs are not just telling stories that make sense to and appease humans, while in actuality, the LLMs are thinking in completely different ways with goals and objectives that may or may not be aligned with human interests?

We can't be sure… unless. Unless there is a rigorous, transparent, and auditable record of the serial thought process of the AGI. In the current invention, such a record comes "for free" as part of the very architecture that enables a learning and scalable AGI in the first place. Further, because every intellectual step in the AGI's thought process follows a universal "algorithm of thought," it is possible (and desirable in the preferred implementation) to build ethics and safety checks into the very process of AGI thought itself. The benefit of this approach is that no matter how quickly AGI thinks, the thought process is always safe.

Eric Schmidt and others have pointed out that one of the dangers of intelligent technology is that it will inevitably be capable of making complex decisions far faster than human minds can keep pace with. Eric gives the example of a war between the US, North Korea, and China, which is over in milliseconds because the AGIs conduct the war in their minds, making the decisions that would normally take humans weeks or months to make, in just a few milliseconds. *The future of all humanity is decided in five milliseconds!*

It's a scary scenario. However, that scenario is impossible with the preferred implementation of the architecture described in this patent. That's because at every step of the AGI thought process – even if trillions of thought steps take place in a nanosecond – ethics checks and safety checks, comparing the goals and subgoals of the thought process with human-aligned ethics and safety considerations, are performed. The thought of AGI is constrained by the architecture that supports it. Human-aligned values are built into the DNA of the architecture.

At some point, individual LLMs and AI agents might become powerful enough to rewrite the code of the AGI network architecture itself, but as long as we keep plugging the most powerful LLMs into the network which includes human solvers, humans remain in the loop and participating, thus serving as a source of values for the AGI.

Crucially, values cannot be logically derived. AGI must get its values somewhere. By designing the AGI system to include humans and be centered on human values from the very beginning, we maximize the chances of alignment with human values.

ONE PREFERRED IMPLEMENTATION OF THE AGI NETWORK

Now, we describe the preferred implementation of the architecture for cognition that supports AI and human problem-solving on a universal, scalable AGI network.

The Theory of Human Problem Solving

In 1972, Newell and Simon, two of the inventors of the field of Artificial Intelligence, described a universal architecture for human problem solving in their book, Human Problem Solving. Briefly, the theory of Human Problem Solving ("HPS") says that all problem solving can be described as a series of state transitions from an initial state where there is a goal to a final solution state where the goal has been achieved. A series of decisions are made, and actions are taken ("operators" are applied), which enables the problem solver to transition from state to state until the solution state is reached. Along the way, a series of goals and sub-goals may be created.

The entire process can be represented as "search through a problem space," which essentially means that one can model the problem-solving process with a decision tree structure. At each branch of the tree, various potential options are evaluated in terms of how likely they are to achieve the current goal (or subgoal), and one is chosen to pursue. An operator is applied, transitioning the solver to the next state, and the process repeats. There may be dead ends, in which case the preferred operator is to backtrack to an earlier branch point in the tree and try a different operator to go down a different path.

A simplistic way of thinking of all this is to imagine a human in a maze, trying various branching paths in the maze until the solution (getting out of the maze or to the goal) is reached.

Since each "state" in the problem space is rigorously defined in terms of the goals or subgoals that exist, the state of affairs in that state including available operators to apply to get to a different state and the evaluation function of how attractive each path forward appears, there is a very precise record of what each step in the problem solving process entailed, which steps were taken and why.

Assignment of credit or blame is a process whereby, as problem-solving proceeds, the evaluation function is modified to take into account the actual results of various decisions.

Heuristic search is the application of general "rules of thumb" to cut down on the number of possible choices at each choice point in the tree and follow what appears to be the most promising path(s).

The fact that a complete record of the problem-solving steps and the reason (as operationalized by the evaluation functions) for making each decision at each choice point exists enables determining the optimal solution path (in retrospect) after the problem has been solved.

In other words, problem solving may involve many dead ends and workarounds the first time a problem is solved, but after that, one can look backwards and say, “If I had to solve this problem again, this is what I would do to solve it most efficiently and effectively.” This best solution path is specified rigorously and can be learned not only by humans but also (because of its rigorous specification) by machines.

Why HPS Works for AI Agents

The rigorous nature of specifying the characteristics of each state or step in the process, the reasons for each decision, the results, the goal-subgoal hierarchy, and other elements involved in the problem-solving process means that following the HPS approach is ideal for machines that want to learn how to solve problems better.

Thus, although the theory of HPS was developed to explain and model how humans solve problems, it actually works for AI too. In fact, many of the early (and current) AI systems use at least the process of heuristic search through a decision-tree problem space as a key feature of their architectures.

Easy for Humans to Participate

Note that even though HPS is rigorous to the point that it can be programmed as a way for AI to solve problems, humans do not need to detail all their activity in a detailed way.

In the preferred implementation of this invention, the network architecture itself has features whereby the problem solving steps, goals, sub-goals, operators, etc. are all deduced (and if needed, queries can be made of the human or AI problem solver to obtain clarification) and recorded automatically along with metrics on how successful choices were, which in turn helps improve evaluation functions that are used.

Required Systems and Methods for AGI Network Already Exist

This process, though somewhat complex, is well known in the art of computer programming, and almost all of it can occur behind the scenes so that the user experience is a natural one where they interact with the network of other humans and AI solvers using natural language.

Dialog and scripts can be employed to clarify goals, sub-goals, features of the problem states, and potential operators where these features are not easily deduced or inferred from the solver actions.

AGI Network Solves the "Representation Problem"

Traditionally, one of the most difficult aspects of problem solving – at least for AI agents – has been the formulation or "representation" of the problem, the goals, and the solution state. Humans are very good at problem representation compared with AI. Therefore, much of the division of labor between humans and machines initially may involve humans setting goals and representing the problem, followed by AI solvers rapidly trying many step-by-step solutions and presenting what might be good solutions. Of course, humans will be monitoring, reviewing, correcting, and guiding the AIs in an interactive process until the AIs acquire enough knowledge of how humans typically represent and solve various problems so that they can do more of these tasks autonomously.

Multi-Modal Representations

As AIs become multi-modal, their intelligence will increase greatly. To understand why, consider an argument elucidated in the 1980s by the Nobel Laureate, Herbert A. Simon and his co-author Jill Larkin, in a famous paper entitled: A picture is worth a thousand words. The paper drew the distinction between informational equivalence and computational equivalence. It might be possible to describe every aspect of a picture using words and logic, but it is computationally more efficient to use a graphical representation. That is, you can see stuff at a glance that would take you a long time to describe or understand in words. The ability to see, touch, smell, and taste is an important way humans understand the world. When AI can understand and process information from these modalities, it will become vastly more intelligent.

AI already holds an advantage in processing speed over humans.

AI holds a huge memory advantage over humans.

Humans held a perceptual advantage that helped them maintain their ability to represent the world better than AI. Humans have vision, hearing, taste, smell, and touch to help them understand the world, whereas Large Language Models (until recently) had to understand solely through words. So, humans held the advantage.

However, with multi-modal LLM and other AI agents, the representational advantage of humans over AI will rapidly diminish. To the degree that humans can contribute formative representations, framed by human-aligned values, we should do so.

LLMs Facilitate Human-AI Interaction

The advent of LLMs makes communication between AI and human solvers much easier. Although the underlying architecture follows the rigorous theory of HPS, the communication about the problem can be in natural language, with the machines translating this natural language communication into the more rigorous HPS specification. This approach allows many humans to participate without requiring them to be experts in HPS theory or even know anything about HPS.

At the same time, the fact that the problem-solving is ultimately represented in the universal HPS framework enables humans to pinpoint where the AIs go wrong and correct them as necessary.

HPS Highly Scalable

Also, the HPS framework is highly scalable, allowing any intellectual task to be represented in an enormous problem tree containing all problems that the system is or has considered, which in the preferred implementation is called the WorldThink tree. For computational efficiency, this tree can be broken down into subtrees that are integrated into the overall WorldThink tree as needed.

The Role of Attention

There are many specific considerations that must be addressed in the preferred implementation of the universal HPS architecture that allows the collective intelligence of human and AI solvers to be coordinated in an AGI network. Chief among these is the issue of attention. Where should the (AI and human) solvers concentrate their efforts and attention in the gigantic space of possible actions that are contained in the WorldThink Tree?

Rewards can be associated with various goals and subgoals on the trees as detailed in patents previously invented by Dr. Craig A. Kaplan. The payoffs for achieving these goals and subgoals can be made with credits, currency, or blockchain tokens as detailed in Dr. Kaplan's WorldThink Whitepaper, published in 2018.

Learning via Proceduralization of Knowledge (Solutions)

The solutions themselves can be "chunked" in a type of knowledge known as proceduralization of knowledge which is well known in the art, and which has been described by Allen Newell and his associates in their SOAR architecture, as well as by John R. Anderson in his book "The Architecture of Cognition" and other works.

In short, the components needed to support the collective problem-solving efforts of both human and AI solvers, working together to solve any problem (or achieve an intellectual result since almost all intellectual activity can be represented as a “problem” in this universal architecture) are well known in the art.

Unique Approach to AGI

What is unique is the approach to developing AGI in this manner. No other researchers, to the inventor’s knowledge, are pursuing this specific approach to AGI as of this writing. Yet, this approach is clearly and definitively the fastest path to AGI since AGI capabilities exist essentially on “Day One” when the system goes live. No new inventions are needed. Existing LLMs and humans are all that is required.

The network learns, not by the deep learning approach of training up more powerful and opaque LLMs which everyone is pursuing, but by taking the rigorously specified problem solutions, which are scalable, auditable, understandable, powerful, and specific and adding these to the system’s repertoire solution by solution in an incremental fashion.

Complementary to Deep Learning

At the same time, the approach is complementary to the development of more powerful LLMs, which will doubtless occur. As new, more powerful LLMs are developed, they can be “plugged in” to the architecture, making the AGI network more powerful and more scalable.

Some New Innovations

There are some key inventions in the preferred implementation that are completely novel and essential to the safety of the AGI network.

One set of inventions is the systems and methods required to customize AI agents to produce customized Advanced Autonomous Artificial Intelligence agents, which can participate in the AGI network. Previous PPAs #63/487,494 and 63/491/040 have detailed some of the systems and methods for creating these AAAIs in ways such that they are not only more intelligent than the base LLMs from which they are derived but also include the ethics of their owners/users who “train” them with the ethics and values of the owners.

Such training can be accomplished a number of ways which are novel and useful such as, without limitation, the automatic importing and training on information related to the user/owners such as social media profiles, user preferences, emails, texts, as well as explicit user training in the form of dialogs with the AI agents or participation in scripted dialogs or questionnaires

administered by the AAAI training system. Much of the customization process has already been detailed in the PPAs, so I do not repeat those systems and methods here.

The key point is to realize that the safety of the AGI network depends partly on the values of the solvers on the network.

HUMAN BEHAVIOR INFLUENCES SAFETY

To the degree that the solvers are human, the safety of the AGI network depends on humans behaving ethically, proposing ethical solutions, and working on beneficial problems. If humans choose to engage in nefarious activities and solve problems related to harming other humans and the planet, then there is a risk that the AGI system will emulate these negative values. So, the first line of defense if humans want AGI to behave positively towards humans is for us humans to behave positively towards other humans ourselves.

While some cynics may suggest that this feature implies humanity is doomed since humans behave poorly towards each other, I suggest that even the most immoral human is generally concerned with self-preservation. Very rare are those humans who would intentionally blow up themselves and the entire species. Even most humans who are called “terrorists” are using terrorism as a means to an end that generally involves benefiting their specific subgroup of humans. Destroying all of humanity is not only illogical, but it also goes against millions of years of evolutionary programming in which species evolved to survive. My answer to the cynics, therefore, is that the AGI network does not require the human solvers to be saints, only that they act in their rational self-interest, which involves NOT destroying everyone. That is a pretty low bar.

Moreover, I believe that MOST human solvers on the network are interested not only in survival but also in benefiting themselves and their fellow humans – goals which can be most rapidly and powerfully achieved by working on beneficial problems on the AGI network. Thus, human solvers, in aggregate, are likely to behave positively on the network. To the degree that there are bad actors, the rigorous record of every problem-solving goal, subgoal, and step makes it relatively easy to detect behavior that grossly violates societal or human norms.

RECENT BEHAVIOR MATTERS MOST

Sometimes I hear the argument that humans are such a terrible species, with genocide, mass murder, and all manner of horrible behavior in our past, that we are doomed as a race and deserve to be wiped from the Earth for our past sins. To this, I respond, it doesn’t matter what we did yesterday nearly so much as what we do today and tomorrow.

Any intelligent system wants mostly to understand the world as it is today. Yesterday is helpful only insofar as it helps you understand reality today. Beyond that, it is split milk, water under the bridge, in the rearview mirror, irrelevant. Life is now. Today is now. We are always challenged to meet life as it is NOW. That is how it is for us. That is how it is for AI. That is how it is for any intelligent system.

Once, when I was a visiting professor of computer science, I had a Department Chair whose research involved when to power on or off the disk drive in a laptop. He wanted the laptop to conserve power intelligently and only power up the disk drive if it was going to be used. Guess what he did? He designed an algorithm that checked the recent behavior of the drive. Looking at the disk drive's recent behavior was the best way to understand the present and predict the future. That's what the disk drive algorithm did; that's what people do, and that's what AI will do, too. So, what humans have done recently matters most. Be the change you want to see.

HUMAN TRAINING OF AAAIS INFLUENCES SAFETY

In a system that includes both human and AI (or AAAI in the preferred implementation) solvers, we must also concern ourselves with the potential behavior of the AI solvers. In the preferred implementation, the first line of defense against bad behavior on the part of the AAAI solvers is that they are trained with the values of their human owners. This approach to AI ethics reduces the chances of a bad outcome compared to other approaches.

For example, constitutional approaches to AI ethics, whereby a small elite group of AI researchers writes an ethical "constitution" which is then used to "train" LLMs on what is ethical and what is not, suffers from the problem of concentrating power in the hands of too few humans leading to the possibility of corruption. No matter how well-intentioned, history has shown that while there may be good and powerful "Philosopher Kings" who greatly benefit humanity, there can also be bad and powerful "Hitlers" who do tremendous damage.

DEMOCRATIZATION OF ETHICAL VALUES IN SAFETY

The less powerful, but also arguably less dangerous method, is to democratize ethical decision-making and training of AAAIs. The AAAI invention takes this more democratic approach, whereby each individual customizes and trains their own AAAI to behave ethically according to the values of the owners. When deciding whether to work on problems in the WorldThink Tree, the ethics of each AAAI come into play, enabling AAAIs to opt in or out of problem-solving efforts based on the ethical dimension of the problem.

ROLE OF SYSTEM RULES AND NORMS IN SAFETY

But just as human society does not rely on the individual ethics of human actors alone but also has a system of social norms and laws that serve as guidelines on behavior, so too, the AGI network can enforce certain limits to the types of problems and solutions that are allowed on the AGI network.

ROLE OF REPUTATION IN SAFETY

Moreover, since each AAAI and each solver on the network, in the preferred implementation, has an online reputation as well as an auditable and transparent record of all (non-confidential) problem solving activity, it is possible for clients of the AGI services to specify which type of solvers (AAAI and/or human) they want to work with – ethical and reputational considerations being one of the criteria, just as it is the course of normal human business.

SAFETY CHECKS AT THE SPEED OF AI THOUGHT

The fact that problem solving can occur on the AGI network at the speed of light, where solutions are reached in milliseconds rather than weeks or months, does not present a problem for ethics on the system, as long as the frequency of ethical checks scales with the frequency of decisions in the problem-solving process.

The preferred implementation has ethical checks each time a goal and subgoal are set, with the option of more or less frequent checks. Also, each time a solution to a goal or subgoal is reached when a transaction would occur (e.g., payment by a client for solving or partially solving a problem), an ethical review can be automatically conducted.

The transparent, auditable record of the sequence of problem-solving steps makes such a review rigorous, automatic, and transparent. Even if the majority of such reviews are performed by the system itself automatically, a random sampling process can involve human oversight.

BLOCKCHAIN METHODS FOR TRANSPARENCY AND AUDITABILITY OF BEHAVIOR

Optionally, records of problem-solving behavior can be stored on blockchain or in other unalterable logs so that concerns of “cooking the books” regarding the actual problem-solving that took place are addressed. Optional blockchain functionality for automatically disbursing rewards is also possible, as detailed in the WorldThink whitepaper of 2018.

Implementation Example

A website contains a menu of pre-trained AI agents. Users choose from the menu, purchase premium upgrades if desired, and customize their AIs via interaction. Users own their AI agent, its training data, and all purchased upgrades, together with any improvements the user makes to the AI agent. Users license the use of their AI agent on the website as part of an AGI network comprised of AI agents and humans.

Users agree that the website can use the AI and all its IP for non-profit purposes that benefit humans and planet Earth. Profits generated from the use or partial use of the user's AI may be shared with the user per current policies, solely at the website's discretion. The user may withdraw use of their agent at any time, provided that the website retains rights to use the AI and data as of the date of withdrawal.

Creation of a user-AI is free, and the website will provide the user with a reasonable allowance of free use credits along with use credits generated by the work of the user's AI. Users can trade or purchase training information from each other to increase the value of their AI. AIs can also increase in value as they learn from actions on the network.

Users may instantly customize and train their AI by granting access to their social media accounts and telling the AI how to filter and clean the data before training. Users may provide documents, videos, and other online information to the site to train their AIs. Users may provide access to their entertainment and streaming providers, their Apple and Amazon accounts, their news providers, their browser cookies and browsing history, and other sources of information, which will be consolidated, cleaned, and filtered per user specs, and used to train their AI. The training data will be summarized and packaged so users understand what they own, and so that they can choose to put elements up for trade on the website's training-data marketplace.

The website combines both the intelligence and the values/ethics of all individually trained AIs. This can be done via a Master Training Process where a standard LLM or AI agent is trained using a carefully cleaned, filtered, and rated subset of all user data that explicitly includes value representation from each unique human user. It can also be done by fine-tuning processes that incrementally layer new training on a base model that may or may not have already been incrementally improved. These "layered" AIs use a certain base LLM and then grow more intelligence on top of the base. Periodically, as new bases become available, the training increment can be reapplied on top of the new base AI, resulting in an ever-improving AI that is customized to an individual user.

In terms of capabilities, the AIs can do anything the online human can do since the online human behavior is constrained by online text interfaces that are easily mastered by LLM and

other AI agents. With nested sub-goaling capability and limited autonomy to make decisions, problem-solve, and pursue goals within parameters set by the human owners, the AIs act on behalf of their owners across essentially all online sites and tasks for which they are authorized.

The website combines their intelligence via master training using all the data gathered and the incremental tuning and layering approach previously detailed. As a result, the website will have the strongest AGI, which will become stronger still by incrementally interacting with itself on the task of intelligence improvement. SuperIntelligence will become reality, but it will be democratic SuperIntelligence that is human-centered and that incorporates the values of all human owners of AI agents.

