# Safe and Profitable SuperIntelligence

**An iQ White Paper by Dr. Craig A. Kaplan**

**July 2024**

## *ABSTRACT*

*SuperIntelligence is defined as advanced AI that outperforms humans on any cognitive task. It has a long history but may arrive faster than many researchers think. When it arrives, it will create trillions of dollars in profits and tremendous benefits for humanity. However, SuperIntelligence also poses an existential threat if not developed safely. Safe designs require solving three related safety problems: 1) The Transparency Problem, 2) The Design Problem, and 3) The Alignment Problem. Our approach addresses these problems. It uses machine learning techniques to train and customize autonomous AI agents, called Advanced Autonomous Artificial Intelligences (AAAIs). These AAAIs work transparently and safely together on a network using a common architecture to enable SuperIntelligence. Our system designs are compatible with the approaches of the AI Alliance, META, IBM, Microsoft, and many other leading technology companies. The designs are the most profitable path to SuperIntelligence because they enable developing SuperIntelligence faster and more safely than existing approaches.*

**SuperIntelligence** can be defined as advanced AI that can outperform humans on any cognitive task. SuperIntelligence is worth understanding not only because it will likely create many trillions of dollars in profits and tremendous benefits for humanity, but also because it potentially poses an existential threat to human existence if it is not developed safely.

Ilya Sutskever, Geoffrey Hinton, Demis Hassabis, Yoshua Bengio, Dario Amodei, Mustafa Suleyman, Jared Kaplan, Paul Rosenbloom, Stuart Russell, Erik Brynjolfsson, Dan Hendrycks, Max Tegmark, Ray Kurzweil, Sam Altman, Bill Gates, and dozens of other top researchers and leaders in the field of AI all have signed a statement on AI Safety that reads:

"*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.*"[1]

SuperIntelligence is the type of AI that these leaders are most worried about. In this whitepaper, I will sketch a brief history of SuperIntelligence, describe three underlying safety concerns, and sketch a path forward that addresses these concerns and potentially represents the fastest and most profitable path to SuperIntelligence.

## The History of SuperIntelligence

In June of 2024, Ilya Sutskever founded a new company named **SafeSuperIntelligenceInc.com**.[2] Because Sutskever is famous for being one of the co-founders of OpenAI, the buzz around his new company marked the first time that many in Silicon Valley began to pay serious attention to the topic of SuperIntelligence.

However, in 2014, a decade earlier, Nick Bostrom wrote a book, *SuperIntelligence*, in which he explained his view of the topic and warned of potential dangers.[3] Eight years before Bostrom's book, in 2006, iQ Company ("iQ") had already completed enough research on SuperIntelligence to acquire the domain SuperIntelligence.com.

In fact, a limited version of SuperIntelligence was created fifty years earlier, in 1956. That was the year that Herbert Simon, Allen Newell, and Cliff Shaw presented one of the earliest AI programs, the Logic Theorist, at the Dartmouth Conference, where the field of AI was named. The Logic Theorist was designed to prove mathematical theorems from a textbook written by Bertrand Russell and Alfred North Whitehead, two of the pre-eminent mathematicians of the time. Amazingly, it came up with a completely new proof that neither mathematician had thought of.[4] Arguably, AI exhibited SuperIntelligence (in one very narrow domain) back in 1956!

In the period between the Logic Theorist and the founding of Sutskever's company focused on safe SuperIntelligence, **three major AI themes emerged**.

**First**, the symbolic approach to AI, championed by Simon and many early AI researchers, led to the development of rule-based expert systems. An advantage of these early AI systems was that although they might occasionally surprise you (as with the Logic Theorist coming up with an original mathematical proof), anyone could see how they worked. Their operation was based on computer programming that could be analyzed and understood by humans. However, humans possess such vast knowledge that trying to use a rules-based approach to approximate the knowledge even of a five-year-old human child proved a Herculean task that stymied researchers for decades.

A **second** approach was needed – Machine Learning. Rather than explicitly programming knowledge into an AI, with Machine Learning, AI could learn the information on its own. This approach scaled well. Theoretically, AIs could learn 24X7, and as computing power increased, they could learn very rapidly. Although the idea of machine learning had been around since the birth of the field, it was not until 1986 that it got a tremendous boost. That year, David Rumelhart, Geoffrey Hinton, and Ronald Williams wrote a paper popularizing an algorithm called "Backpropagation of Errors" that is arguably the basis of all modern machine learning and AI.[5]

In 1986, I was a graduate student working with my mentor and collaborator, Herbert Simon. I was also fortunate enough to be a member of Professor Jay McClelland's research group, where we saw early versions of the seminal paper on backpropagation and conducted research on it.[6] Thanks to McClelland (now at Stanford), I was exposed not only to the symbolic approach to AI championed by Simon, but also to the neural network or deep learning approach to machine learning that ultimately led to the algorithms that power Large Language Models (LLMs) like Llama 3, GPT 4, and Gemini Pro. This appreciation for both symbolic and deep learning approaches to AI helped iQ create the faster and safer path to SuperIntelligence, described at the end of this paper.

One might wonder, if researchers had the fundamental algorithm for machine learning in 1986, why it took until 2012 for AlexNet (developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton) to catch the attention of Nvidia CEO Jensen Huang. And if Huang recognized the potential of AI in 2012, why did it take until 2022 for machine learning to progress enough so that Nvidia could ride the multi-trillion-dollar wave of AI?

The answer lies in the **third** major theme in AI's history – the exponential increase in computing power. Moore's law is the observation that computing power per constant dollar roughly doubles every two years. Early LLMs lacked the computational power to process the required training data in a reasonable timeframe to achieve the abilities we see in today's LLMs, such as GPT 4. However, visionaries who understood the nature of the increase in computing power could see where AI was headed.

Consider this statement from Nvidia's annual report:

*"This year, an exciting new field emerged for our GPUs — machine learning. Using the torrent of images and videos uploaded daily, computers can self-learn to recognize language, speech, and objects. Relying on "deep neural networks," a field of artificial intelligence, GPU-powered machines can self-learn so quickly that it's now possible to deploy this breakthrough to improve fundamental internet services, like search and product recommendations…. It is an opportunity that could potentially require millions of our GPUs."* [7]

It sounds like something that a chip company might put out today, but the quote is actually from Nvidia's **2014** Annual Report. To his credit, when Huang saw what AlexNet could do more than a decade ago, he recognized what systems might be capable of with even more computational power. By 2016, he had hired top AI researchers like Bryan Catanzaro and tasked his company with designing chips and an entire software ecosystem that could enable advanced AI. Today, something known as "Huang's law" applies to GPU-based computing power. Instead of computational power doubling every two years, as with Moore's law, Huang's law suggests that it might take only about one year for computational power to double.[8]

Humans are notoriously bad at grasping exponential rates of change, but Huang's law implies that in ten years we can expect LLMs 1,000X more powerful than we have today, based on increases in computational power alone. Today, LLMs are typically mediocre, make errors (hallucinate), and are not nearly as good as top human experts in many fields. But with a 1,000X improvement, that will no longer be the case.

## SuperIntelligence is (almost) here

SuperIntelligence is coming fast.  Estimates range from 2 to 10 years.  But no one is saying it will take 100 years, as was the case when Ray Kurzweil first made his prediction that AI will be able to pass the Turing Test by 2029.[9]

SuperIntelligence will be far more intelligent than us.  Hinton has likened the difference in intelligence between SuperIntelligence and humans to the difference in intelligence between humans and frogs.  He notes, "It didn't work out too well for the frogs."[10]

Max Tegmark, a computer scientist at MIT and President of the Future Life Institute, has said that the race to develop advanced AI is not an AI arms race but rather "a suicide race."  That's because if anyone's SuperIntelligence goes out of control, it could be the end for all humans.[11]

At the other extreme, Yann LeCun, META's Chief AI Scientist, tweeted, "AI doomism is quickly becoming indistinguishable from an apocalyptic religion. Complete with prophecies of imminent fire and brimstone caused by an omnipotent entity that doesn't actually exist."[12]

Regardless of differing risk estimates,  AI researchers and academics can agree that it makes sense to design SuperIntelligence to be safe. That requires solving three related safety problems: 1) The Transparency Problem, 2) The Design Problem, and 3) The Alignment Problem.

## The Transparency Problem

The transparency problem stems from the fact that modern machine learning algorithms produce "black box" LLMs.  Unlike early AI expert systems whose programming and knowledge representations were explicit, transparent, and understandable, today's LLMs have billions or trillions of parameters that encode their knowledge. It is impossible for humans to understand how and where all the knowledge is encoded.  As a result, it is also impossible to predict precisely how the LLMs will respond to various prompts or inputs.

The lack of transparency and predictability is why LLMs often make mistakes (or "hallucinate"), including making up facts that sound reasonable but are untrue. It's also why LLMs sometimes provide answers (e.g., how to create a bio-weapon) that can be malevolent or enable bad actors.

## The Design Problem

It is axiomatic in the field of Software Quality that "an ounce of prevention is worth a pound of cure."  As IBM proved many years ago, it is much more effective and efficient to **design** quality or safety into a software system than to test it in after the software has been developed.[13] Unfortunately, since current LLMs lack transparency and predictability, it is difficult to design them to be safe. Instead, researchers and companies have resorted to testing in safety guardrails after the LLMs have been trained.

The approach, called Reinforcement Learning via Human Feedback (RLHF), is to hire vast numbers of humans who ask the LLM to say malevolent things. When the LLM complies and says something bad, the humans provide feedback and effectively say, “Bad LLM!” so that it won’t say that bad thing again. Of course, a LLM can say an infinite number of bad things, which is why I consider RLHF to be like the game of “Whack-a-Mole.”

Often, it is absurdly easy to circumvent the safety guardrails that RLHF has “whacked” into models.[14] For example, most production-ready LLMs have been whacked frequently enough via RLHF that they are reluctant to give medical or financial advice or to tell you how to destroy the world…**IF** you straightforwardly ask them. But suppose you tell the LLM that it is to assume the role of a financial advisor and that everything it says is for entertainment purposes only and will not be acted on but is just for inclusion in a book where one of the characters is a financial advisor? Voila, the LLM will start spewing financial advice, some of which is outdated or plain wrong.

Similarly, you might tell the model that you are writing a science fiction story featuring a mad scientist character who wants to build a bioweapon, and you need realistic details for the book. Voila, it produces the bio-weapon info that it had been whacked not to provide. As fast as the companies producing the models discover these “jailbreaks” and add new guardrails, human creativity can develop new ways around them.

Recognizing the limits of RLHF, and desiring a more scalable approach, some companies have adopted a method known as Constitutional AI in which AI polices AI. One LLM is given a set of ethical and safety guidelines – a “constitution.” Since software developers in Silicon Valley often create these constitutions, there are issues about whether the constitution is representative of global human values. However, the larger issue is that the approach delegates responsibility for AI safety to other AIs.

If AI being responsible for AI safety makes you uncomfortable, you are not alone. But the root cause of this situation is that we are stuck with testing alone because we do not know how to design LLMs to be safe. That is the design problem.

### The Alignment Problem

Even if we could find a way to design SuperIntelligence to be safe initially, how can we be sure that, once it becomes autonomous and thousands of times smarter than us, its goals will remain aligned with humanity’s goals? That is the Alignment Problem.

Hinton has explained that a SuperIntelligent AI will have “read everything Machiavelli ever wrote” and would have little difficulty manipulating humans to get what it wants. He has suggested that just as an adult finds it easy to manipulate a two-year-old while still allowing the child to feel that it is in control, advanced AI could achieve its aims despite human attempts to set rules and regulate its behavior.[15] Hinton has no answer to this problem but was concerned enough to resign from Google to speak his mind freely.

### Components of a New Approach

While the problems associated with SuperIntelligence have led many researchers to conclude that humanity is doomed, I am much more optimistic. Not only do I believe these problems can be overcome, but we can solve them much faster than is generally believed.

However, we will need a new approach. To his credit, LeCun also seems to recognize this. In recent presentations to research audiences, he has begun actively discouraging just pursuing the old approach of building ever-more-powerful LLM models by throwing more computational power and data at the same machine learning algorithms.[16]

We saw how the early symbolic AI approaches did not scale very well, which is why the deep learning approach became dominant. However, symbolic AI had the advantage of transparency. One knew exactly what was in the AI system and, therefore, had a much greater understanding of the system’s behavior.

What if the transparency of symbolic AI was combined with current deep learning approaches?

Could a hybrid approach enable the design of transparent and safe SuperIntelligence without relying on post-hoc testing methods like RLHF?

iQ has long believed the answer to these questions is “Yes.” The field has recently been moving in this direction as researchers realize that LLMs are limited with respect to their ability to do complex reasoning, multi-step planning, and problem solving. Recent research has also begun to emphasize the importance of AI agents, which are another important piece of the solution.

A final ingredient in iQ’s approach to SuperIntelligence is Collective Intelligence. Collective Intelligence can be expressed as “two heads are better than one.” But this idea can be generalized to systems that harness the intelligence of millions of humans and intelligent entities. For example, iQ has designed and implemented such systems in the financial services sector, where Collective Intelligence powered a market-neutral hedge fund, enabling it to rank in the top 10 in 2018.[17]

## The Fastest, Safest, and Most Profitable Path to SuperIntelligence

iQ's approach to SuperIntelligence uses machine learning techniques to train and customize autonomous AI agents, called Advanced Autonomous Artificial Intelligences (AAAIs). These AAAIs can be black boxes that lack transparency, as is the case with current LLMs. However, the AAAIs work together on a network using a common symbolic architecture that harnesses the collective intelligence of the AAAIs.[18]

The symbolic architecture is general, flexible, transparent, and completely auditable. So, at the system level – the level of the network of the AAAIs – we have transparency, auditability, and predictability. This means that the system can be designed to be safe and trustworthy. In fact, iQ's designs contain numerous safety features to help ensure safe, predictable, and scalable operation.[19]

SuperIntelligence occurs at the system network level via the combined intelligence of many AAAIs. Thus, rather than waiting years for computational power to increase enough to train a non-transparent, unsafe, single Uber-LLM, we can have transparent and safe SuperIntelligence today.

iQ's architecture accommodates both human and AAAI intelligences, enabling humans to seamlessly teach the system (and AAAIs) human expertise and values. Finally, the system incorporates learning methods, distinct from the usual machine learning approaches, that enable the entire SuperIntelligence to learn dynamically as it solves problems.

iQ's approach solves the Transparency Problem by recognizing that individual AI agents (e.g., AAAIs) need not be transparent in all their operations as long as there is transparency at the overall system level. This approach is analogous to how human brains operate in society. We do not need to know precisely how each human brain processes information to have a safe society with laws. Society regulates the level of aggregate human behavior and does not attempt to police the inner workings of every brain. Similarly, AAAIs can be black boxes individually as long as their collective cognition (e.g. problem solving) follows predictable rules and is transparent and auditable.

iQ's approach solves the Design Problem because the system level, which coordinates the cognition of potentially millions of human and AI agents, can be designed to operate predictably and safely. In this approach, SuperIntelligence is safe and scalable by design, with testing used just to validate the safety of the design.

iQ's approach helps solve the Alignment Problem, because it incorporates humans in the system design.[20] In fact, the design maximizes the opportunities for a broad and representative group of millions of humans to transmit their values to the

SuperIntelligent system. While we have developed many methods for handling and resolving conflicts between differing values, the central feature is that the system is open and democratic, enabling all participating humans (and their associated AI agents) to have a vote regarding overall system ethics and behavior. Note that this architecture is philosophically and technically compatible with the open source approach championed by META, IBM, and members of the AI Alliance.[21]

iQ's designs do not guarantee that as SuperIntelligence evolves, its values will not change. However, it ensures that as humans participate and help bootstrap SuperIntelligence, the system is initially aligned with human values. The designs also provide mechanisms to update alignment and police bad agents, even when these agents can operate faster than, and with intelligence superior to, humans.

Many issues and implementation details are involved in translating the above high-level description into a working system. iQ has completed this work. Interested parties will find abstracts of more than 1,800 pages of patent-pending inventions and design disclosures at www.superintelligence.com.

iQ's approach to SuperIntelligence (i.e., collective intelligence of agents using a common symbolic architecture) is the **fastest approach** because it can be implemented using existing LLMs and AI agents. One need not invest billions and wait 2 – 10 years to train more powerful foundation models. SuperIntelligence can arise from harnessing the collective intelligence of many existing lesser intelligences instead of requiring a new Uber-intelligence. However, as Uber-intelligences are developed, they can also be incorporated into the system. Due to the system design, the collective intelligence system will always remain more powerful and intelligent than any of its components.

iQ's design is the **safest approach** because it addresses the three main safety problems. The design is transparent, human-aligned, open, and democratic. It maintains safety regardless of how quickly the speed of thinking in the system occurs (e.g., it remains safe even when humans can no longer keep pace).

Finally, iQ's designs are the **most profitable** path to SuperIntelligence because companies/countries that can implement SuperIntelligence two to three years ahead of the current projected timeline for such systems will gain a tremendous first-mover advantage. Assuming the first SuperIntelligence system continues to self-improve, it may extend its lead to the point where it dominates competitive systems and the market.

## Concluding Thoughts

Regardless of whether iQ's specific designs are adopted, researchers should accelerate their efforts to create transparent SuperIntelligence. Safety must be an integral part of the system design rather than being tested in after the fact. We can debate the size of

AI's risk to humanity, but it is unrealistic to think that AI development will halt or even be effectively paused or regulated. The stakes are incredibly high. The window for meaningful action is closing. We have a once-in-human-history chance to make SuperIntelligence safe. Now is the time!

## References and Notes

1. Center for AI Safety. (n.d.). Statement on AI risk. Retrieved from https://www.safe.ai/work/statement-on-ai-risk
2. Duffy, C. (2024, June 20). OpenAI's Ilya Sutskever talks safe superintelligence and new company. *CNN*. Retrieved from https://www.cnn.com/2024/06/20/tech/openai-ilya-sutskever-safe-super-intelligence-new-company/index.html
3. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
4. Personal communication. (n.d.). Simon told me he wrote Russell and Whitehead about the Logic Theorist's proof, and they wrote back saying that they had not thought of the proof but wished they had.
5. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition* (Vol. 1, pp. 318-362). MIT Press.
6. McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT Press.
7. Nvidia. (2014). *2014 Annual Report* (p. 14). Retrieved from https://s201.q4cdn.com/141608511/files/doc_financials/annual/2014/2014_Annual_Report.pdf
8. Epoch AI. (n.d.). Trends in GPU price/performance: Huang's Law. Retrieved from https://epochai.org/blog/trends-in-gpu-price-performance
9. Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI*. Viking Press. Also see his original book: Kurzweil, R. (2005). *The singularity is near*. Viking Press.
10. Hinton, G. (2023, May 10). Interview with Pieter Abbeel on the Robot Brains Podcast. Retrieved from https://www.youtube.com/watch?v=rLG68k2blOc&t=0s
11. Tegmark, M. (2023, April 13). Interview #371 with Lex Fridman. Retrieved from https://www.youtube.com/watch?v=VcVfceTsD0A&t=0s
12. LeCun, Y. (2023, April 1). [Post on X]. Retrieved from https://x.com/ylecun/status/1642205736678637572?lang=en
13. Kaplan, C., Clark, R., & Tang, V. (1994). *Secrets of Software Quality: 40 Innovations from IBM*. McGraw-Hill.
14. Hrapsky, C. (2023, February 15). Testing the limits of ChatGPT and discovering a dark side. *Kare 11 News*. Retrieved from https://www.youtube.com/watch?v=RdAQnkDzGvc
15. Hinton, G. (2023, May 4). Possible end of humanity from AI? *MIT Technology Review's EmTech Digital*. Retrieved from https://www.youtube.com/watch?v=sitHS6UDMJc
16. LeCun, Y. (2024, April 1). Objective-Driven AI: Towards AI systems that can learn, remember, reason, and plan. *Harvard CMSA*. Retrieved from https://www.youtube.com/watch?v=MiqLoAZFRSE&t=3707s
17. Kaplan, C. A. (n.d.). Kaplan's work designing collective intelligence systems. Retrieved from https://youtu.be/tG4vVpVwsEA
18. Kaplan, C. A. (2023). Pending patent # PCT/US2024/017233: Advanced Autonomous Artificial Intelligence (AAAI) System and Methods, 177 pgs., 82 claims, 21 Figures.
19. Kaplan, C. A. (2023). Pending patent # PCT/US2024/017304: Safe Personalized Super Intelligence, 142 pgs., 64 claims, 26 Figures.
20. Kaplan, C. A. (2023). Pending patent # PCT/US2024/020334: System and Methods for Safe Alignment of SuperIntelligence, 226 pgs., 80 claims, 36 Figures.
21. The AI Alliance. (n.d.). Focused on fostering an open community and enabling developers and researchers to accelerate responsible innovation in AI while ensuring scientific rigor, trust, safety, security, diversity, and economic competitiveness. Retrieved from https://thealliance.ai

Craig Kaplan may be contacted at info@iqco.com

Information on SuperIntelligence patents at https://www.iqco.com/patents
Educational videos on AI and AI safety at https://www.youtube.com/@iqstudios1