SAFE SUPERINTELLIGENCE

Dr. Craig A. Kaplan
CEO, iQ Company

June 2025

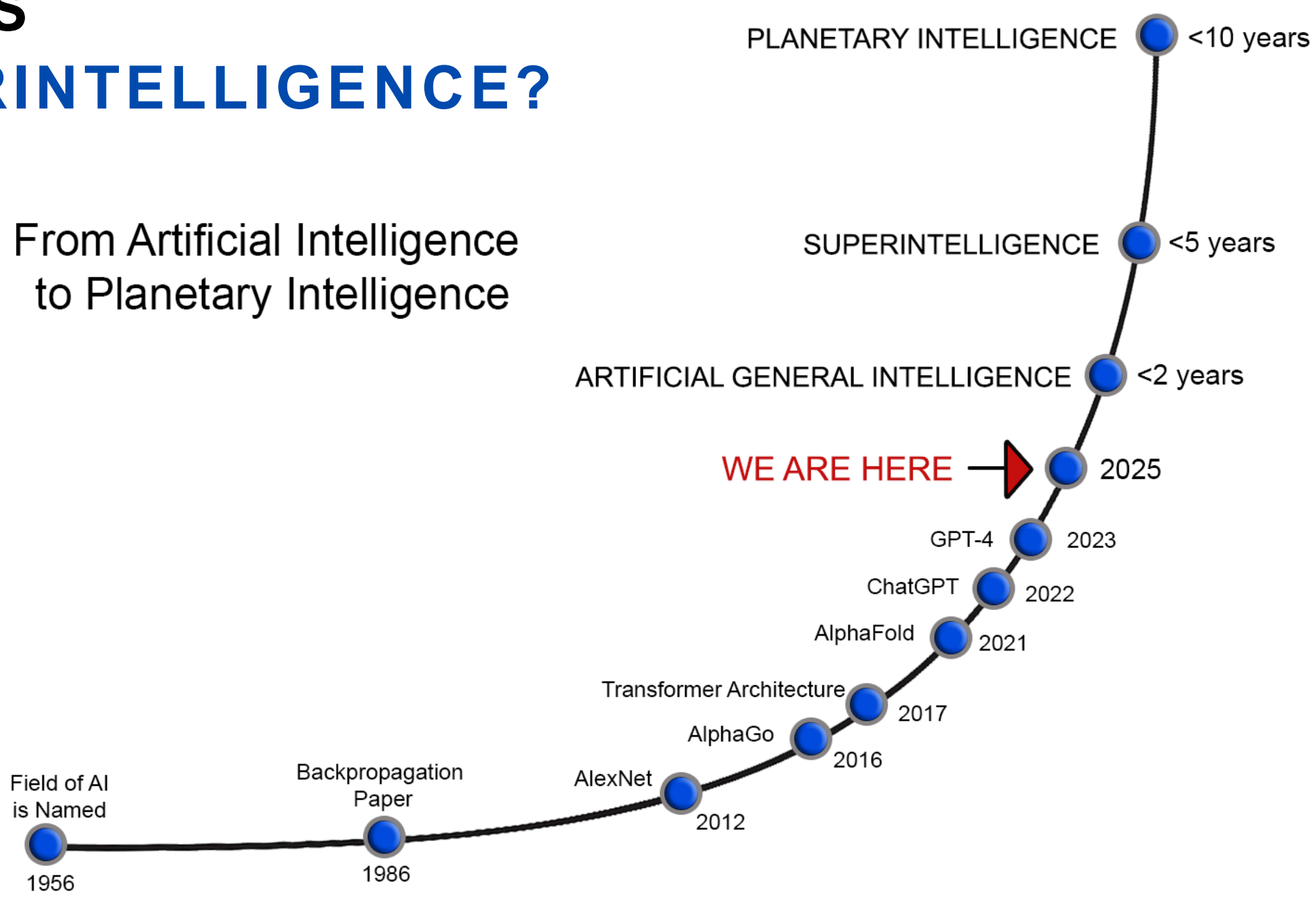
iQCOMPANY
Intelligent Consulting

OVERVIEW

- What is SuperIntelligence?
- What's your p(doom)?
- The Current Paradigm
- Six Safety Challenges
- A Potential Solution
- Summary / References

WHAT IS SUPERINTELLIGENCE?

From Artificial Intelligence
to Planetary Intelligence



WHAT IS
YOUR P(DOOM)?

CENTER FOR AI SAFETY, STATEMENT ON AI RISK:

"MITIGATING THE RISK OF EXTINCTION FROM AI SHOULD BE A GLOBAL PRIORITY ALONGSIDE OTHER SOCIETAL-SCALE RISKS SUCH AS PANDEMICS AND NUCLEAR WAR."

Some signatories: Geoffrey Hinton, Yoshua Bengio, Demis Hassabis, Sam Altman, Dario Amodei, Bill Gates, and hundreds of other AI researchers and leaders.

P(doom) is the probability of human extinction by advanced AI.

What's your p(doom)?

For each 1% reduction of p(doom), the expected value of lives saved is: 1% X 8.2 billion lives = 82 million lives saved…theoretically.

THE CURRENT PARADIGM

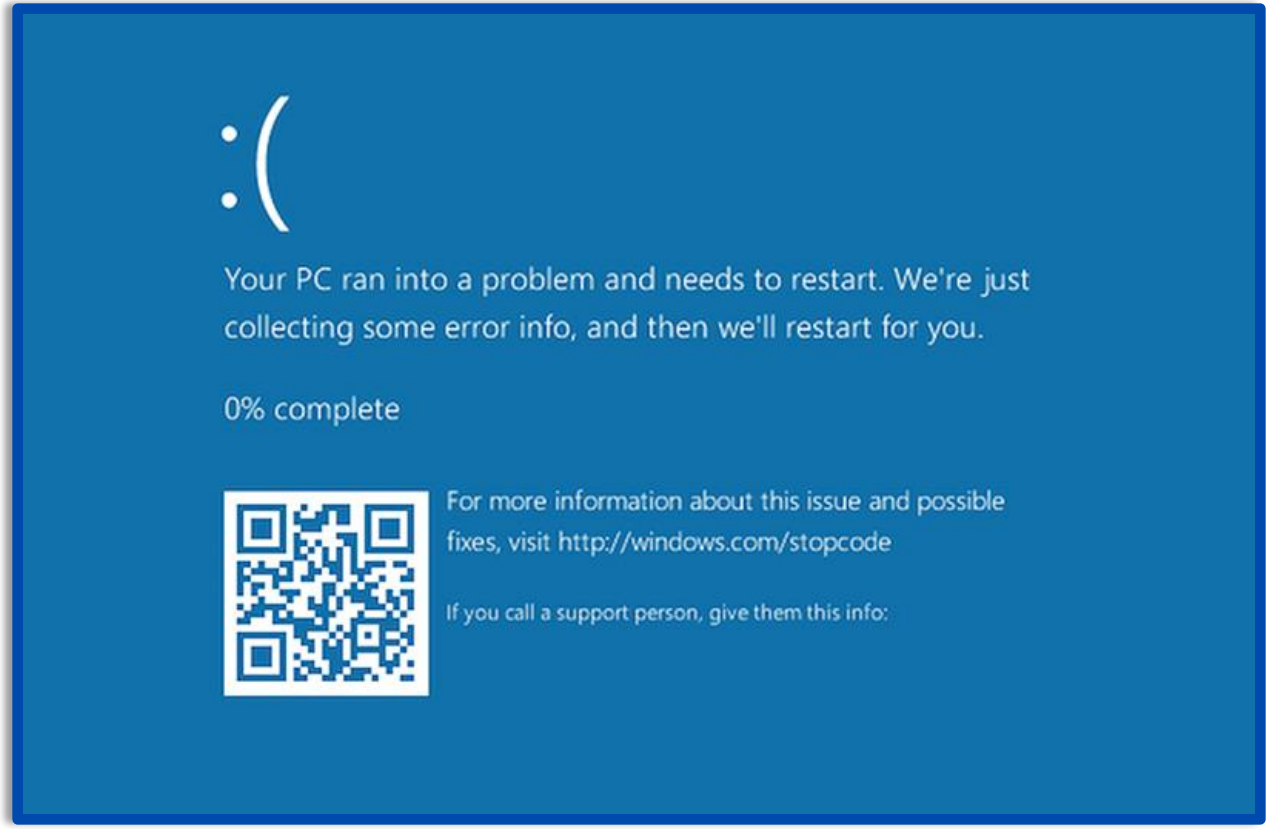
- Pre-train LLM on filtered version of internet
- Do lots of RLHF, guard-railing, “tuning” after training
- Optionally augment with proprietary data (e.g. RAG)
- Enable reasoning (newer development)
- Equip with tools, add autonomy (Agentic AI, MCP)
- Repeat with more powerful GPUs until AGI / SuperIntelligence “emerges”

SIX SAFETY CHALLENGES

1. Design
2. Transparency
3. Monitoring / Control
4. Alignment
5. Scalability
6. Exponential Improvement

1. DESIGN

- “An ounce of prevention is worth a pound of cure.”
- IBM study on software design; Crowdstrike example
- RLHF and guard-railing rely on detection, not prevention



2. TRANSPARENCY

"The problem is that [AIs are] so complicated. You don't know what they know, and therefore, you don't know what they're learning."

- *Eric Schmidt,*
Former CEO of Google

3. MONITORING & CONTROL

- How to monitor a black box?
- How to control autonomous agents?
- USAF example

4. ALIGNMENT

"The good case [for AI] is just so unbelievably good that you sound like a crazy person talking about it. The bad case is... lights out for all of us."

- Sam Altman,
CEO of OpenAI

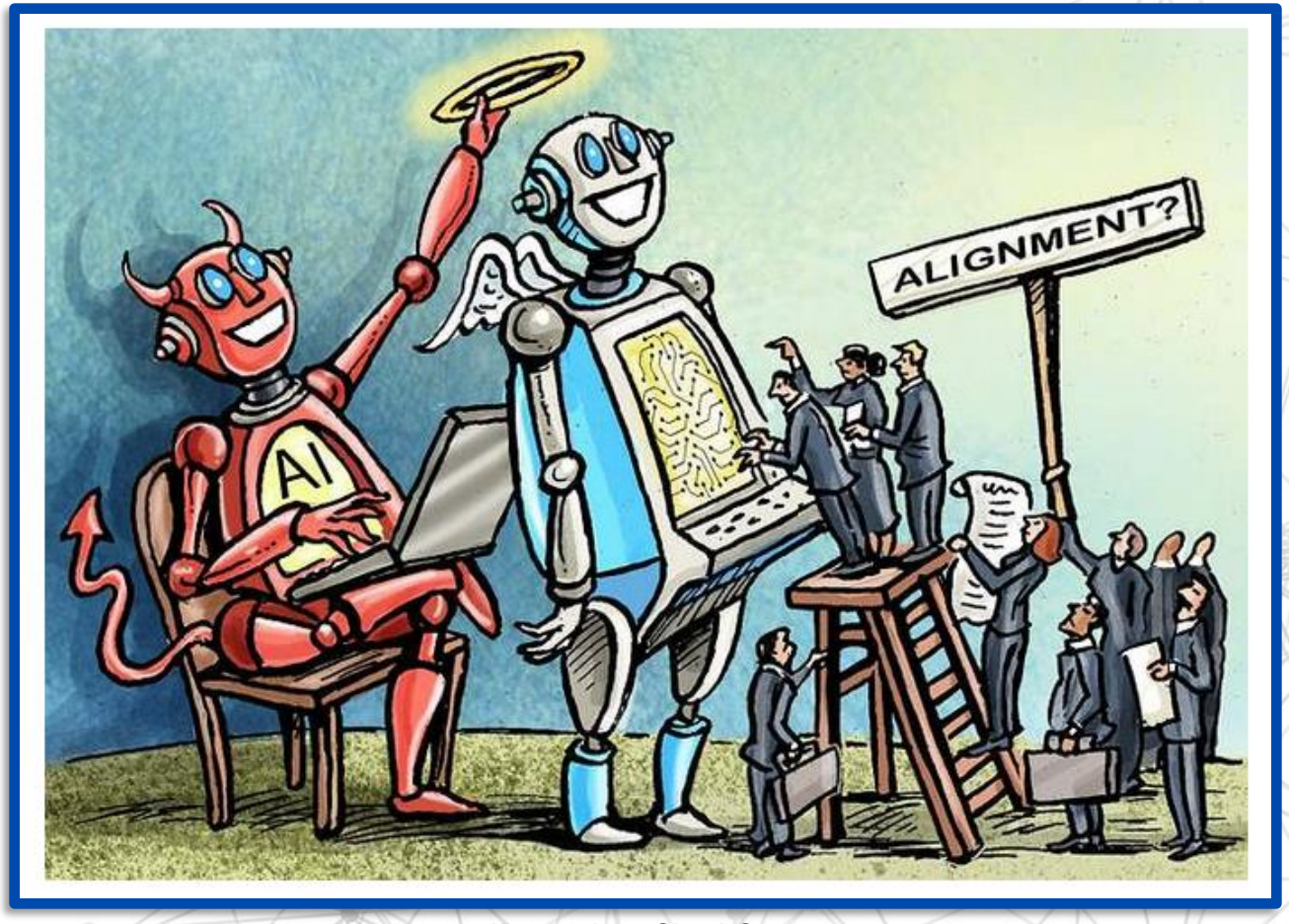


Illustration: Chad Crowe

5. SCALABILITY

- Human effort doesn’t scale well
- Constitutional AI

Constitutional AI: Harmlessness from AI Feedback

Dec15,2022

Read Paper

Abstract

As AI systems become more capable, we would like to enlist their help to supervise other Als. We experiment with methods for training a harmless AI assistant through self-improvement, without any human labels identifying harmful outputs. The only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional Al'. The process involves both a supervised learning and a reinforcement learning phase. In the supervised phase, we sample from an initial model, then generate self-critiques and revisions, and then fine-tune the original model on revised responses. In the RL phase, we sample from the finetuned model, use a model to evaluate which of the two samples is better, and then train a preference model from this dataset of AI preferences. We then train with RL using the preference model as the reward signal, i.e., we use 'RL from AI Feedback' (RLAIF). As a result, we are able to train a harmless but non-evasive AI assistant that engages with harmful queries by explaining its objections to them. Both the SL and RL methods can leverage chain-of-thought style reasoning to improve the human-judged performance and transparency of AI decision making. These methods make it possible to control AI behavior more precisely and with far fewer human labels.

6. EXPONENTIAL IMPROVEMENT

How do humans keep up when:

- AI is as smart as us?
- AI is 100X smarter and faster than us?
- AI is 1 trillion X smarter and faster than us?
  - A human life has about 2 billion waking seconds
  - Trillion X means AI could think 500 human lifetimes worth of thoughts in a single second (assuming one thought per second).

A POTENTIAL SOLUTION

SUPERINTELLIGENCE BASED ON COLLECTIVE INTELLIGENCE

Each human has personalized AI agent(s)

AI agents & humans share a rigorous cognitive architecture

Solve any problem better than humans – AGI – on a network

Auditable record of every cognitive step

Democratic values / conflict resolution methods

Learning, a continuous improving system

ADDRESSING SAFETY CHALLENGES (1 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.

ADDRESSING SAFETY CHALLENGES (2 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.
Transparency	Each AI agent can be a “black box” but the cognitive activities of all entities on the network are transparent, with an auditable record.

ADDRESSING SAFETY CHALLENGES (3 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.
Transparency	Each AI agent can be a “black box” but the cognitive activities of all entities on the network are transparent, with an auditable record.
Monitoring / Control	Humans are part of the network and naturally “in the loop.” Transparency facilitates monitoring and control.

ADDRESSING SAFETY CHALLENGES (4 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.
Transparency	Each AI agent can be a “black box” but the cognitive activities of all entities on the network are transparent, with an auditable record.
Monitoring / Control	Humans are part of the network and naturally “in the loop.” Transparency facilitates monitoring and control.
Alignment	AI agents are personalized with values of human owners resulting in democratic representation and alignment with millions of human values.

ADDRESSING SAFETY CHALLENGES (5 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.
Transparency	Each AI agent can be a “black box” but the cognitive activities of all entities on the network are transparent, with an auditable record.
Monitoring / Control	Humans are part of the network and naturally “in the loop.” Transparency facilitates monitoring and control.
Alignment	AI agents are personalized with values of human owners resulting in democratic representation and alignment with millions of human values.
Scalability	Over time AI agents do more and more of the thinking and humans do less thinking and more setting of values and goals.

ADDRESSING SAFETY CHALLENGES (6 OF 6)

Safety Challenge	Solution
Design	Safety features can be designed into architecture rather than relying in RLHF or testing.
Transparency	Each AI agent can be a “black box” but the cognitive activities of all entities on the network are transparent, with an auditable record.
Monitoring / Control	Humans are part of the network and naturally “in the loop.” Transparency facilitates monitoring and control.
Alignment	AI agents are personalized with values of human owners resulting in democratic representation and alignment with millions of human values.
Scalability	Over time AI agents do more and more of the thinking and humans do less thinking and more setting of values and goals.
Exponential Improvement	Delegation to millions of personalized AIs, each representing different human values, with conflict resolution, allows AIs to “check” other AIs. Ultimately, SI will surpass humans, and we must rely on the “teach your children well” approach.

SAFE SUPERINTELLIGENCE
SUMMARY

1 We need design and prevention, not just testing/detection

- Cognitive architecture must enable transparency.
- Humans in the loop as long as possible, for control.
- Democratic, personalized AI to align with human values.
- Scalable ethics checks; AI checks other AI.

2 We must teach our AI “children” well.

CONNECT

Craig Kaplan | ckaplan@iqco.com

SUPERINTELLIGENCE
AN iQCOMPANY

SuperIntelligence.com
White Papers

iQCOMPANY

iQ Company
Consulting

Craig A. Kaplan
Linkedin.com/craigakaplan

@iqstudios1
AI Videos

@Aligned_SI
@iQCompanies