HOW TO CREATE AGI **AND NOT DIE**

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

June 11, 2024 | San Francisco, California

Dr. Craig A. Kaplan
CEO, iQ Company
ckaplan@iqco.com

WHAT IS AGI?

WHAT IS AGI?

- AGI: Artificial General Intelligence
- Defined as AI with average human ability in any cognitive area

IS AGI DANGEROUS?

"More dangerous than nuclear weapons… by a lot."

-- Elon Musk

"More dangerous than nuclear weapons… by a lot."

-- Elon Musk

"AI doomism is quickly becoming indistinguishable from an apocalyptic religion." -- Yann LeCun

Contents

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

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever

NO PAUSE IN SIGHT

Max Tegmark, MIT & Future of Life Institute

Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33711

Add your signature

Published
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and management is not happening, even though

CHALLENGE #1:
ALIGN SAFETY WITH PROFITS

- Usually, safer implies slower

CHALLENGE #1:
ALIGN SAFETY WITH PROFITS

- Usually, safer implies slower
- Companies/countries fear competitors won’t slow down

CHALLENGE #1:
ALIGN SAFETY WITH PROFITS

- Usually, safer implies slower
- Companies/countries fear competitors won't slow down
- Result: a race to the bottom – Tegmark's "suicide race"

CHALLENGE #1:
ALIGN SAFETY WITH PROFITS

WHAT **IF**:

- The fastest path to AGI was also the safest path?

CHALLENGE #2:
ERROR PREVENTION NOT DETECTION

- An ounce of prevention is worth a pound of cure

CHALLENGE #2:
ERROR PREVENTION NOT DETECTION

- An ounce of prevention is worth a pound of cure
- Lack of transparency → Error detection via testing / RLHF

CHALLENGE #2:
ERROR PREVENTION NOT DETECTION

- An ounce of prevention is worth a pound of cure
- Lack of transparency → Error detection via testing / RLHF
- Scale of testing problem → Constitutional AI / auto-testing

CHALLENGE #2:
ERROR PREVENTION NOT DETECTION

- An ounce of prevention is worth a pound of cure
- Lack of transparency → Error detection via testing / RLHF
- Scale of testing problem → Constitutional AI / auto-testing
- Root causes:
  - We don’t know how LLMs really work / represent info
  - We don’t know how to design transparent, safe AGI

CHALLENGE #2:
ERROR PREVENTION NOT DETECTION

WHAT **IF**:

- We could **design** transparent and safe AGI?
- We designed safety into AGI and used testing (e.g. RLHF) only for verification?

WHAT IS THE FASTEST PATH?

1) Online Network of Humans

WHAT IS THE FASTEST PATH?

1) Online Network of Humans
2) Add Customized AI (“Agents”)

WHAT IS THE FASTEST PATH?

1) Online Network of Humans
2) Add Customized AI (“Agents”)
3) Plug into Universal Cognitive Architecture

WHAT IS THE FASTEST PATH?

1) Online Network of Humans
2) Add Customized AI (“Agents”)
3) Plug into Universal Cognitive Architecture
4) Collective Intelligence of **Human and AI Agents → AGI**

WHY IS THE FASTEST PATH THE SAFEST PATH?

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Existential threat is “the alignment problem”

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Existential threat is “the alignment problem”
- Safe AGI = Aligned AGI

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Existential threat is “the alignment problem”
- Safe AGI = Aligned AGI
- Values cannot be rationally derived, but AGI can learn human values

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Existential threat is “the alignment problem”
- Safe AGI = Aligned AGI
- Values cannot be rationally derived, but AGI can learn human values
- Values could (and should) come from many humans working with AI agents in a collective AGI system

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Customized AI agents include knowledge **and values** of owners

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Customized AI agents include knowledge and values of owners
- AI agents may be “black boxes” but AGI network is transparent and auditable due to features of the cognitive architecture

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Customized AI agents include knowledge and values of owners
- AI agents may be “black boxes” but AGI network is transparent and auditable due to features of the cognitive architecture
- Scalable safety checks / mechanisms can be built into cognitive architecture

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Customized AI agents include knowledge and values of owners
- AI agents may be “black boxes” but AGI network is transparent and auditable due to features of the cognitive architecture
- Scalable safety checks / mechanisms can be built into cognitive architecture
- “Humans in the loop” maximizes opportunity to align AGI system with human values

WHY IS THE FASTEST PATH THE SAFEST PATH?

- Customized AI agents include knowledge and values of owners
- AI agents may be “black boxes” but AGI network is transparent and auditable due to features of the cognitive architecture
- Scalable safety checks / mechanisms can be built into cognitive architecture
- “Humans in the loop” maximizes opportunity to align AGI system with human values
- Include methods to dynamically update and resolve conflicts between values of entities

SUMMARY

- Challenges are: 1) Align safety and profits; 2) Design safety in
- Collective intelligence of humans and customized AI agents using transparent cognitive architecture could meet both challenges
- May be other approaches…collective intelligence of many researchers will be needed to create safe, transparent AGI

LEARN MORE & CONNECT

ckaplan@iqco.com

iqco.com

iqstudios.org

@iQCompanies

@iqstudios1

linkedin.com/in/craigakaplan

@iqstudios_1