Provably Safe AGI

Steve Omohundro
Beneficial AI Research

AGI and ASI are Imminent

- Metaculus “Weak AGI”: March 14, 2026

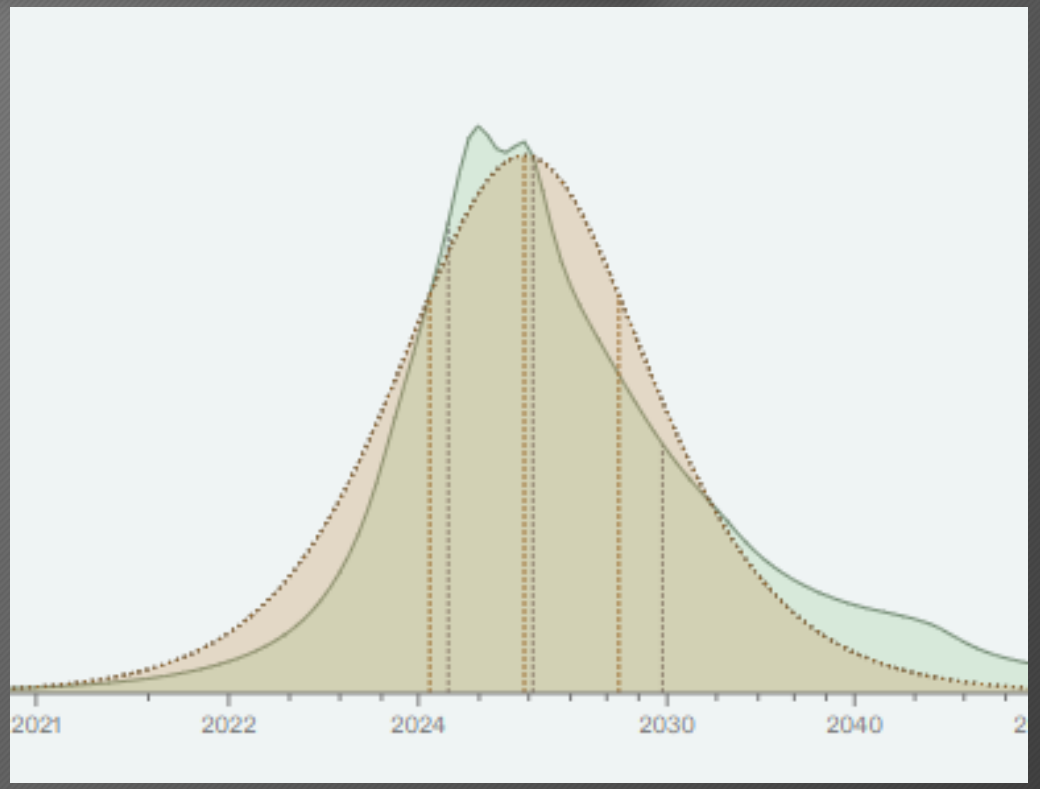
  https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/

- Metaculus “AGI with robots”: May 7, 2031

  https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/

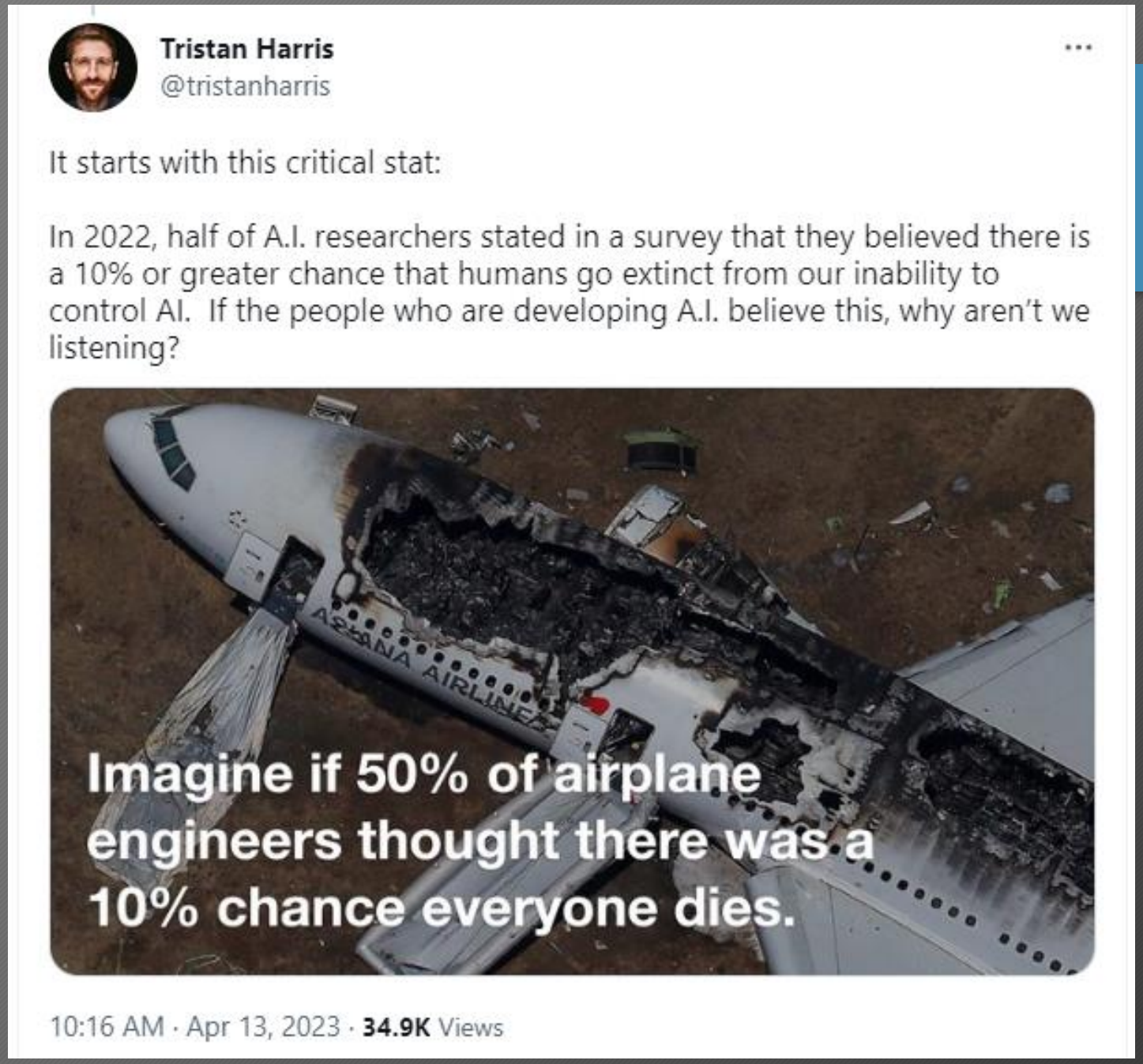
- Metaculus “ASI arrival after AGI”: 6 months

  https://www.metaculus.com/questions/4123/after-an-agi-is-created-how-many-months-will-it-be-before-the-first-superintelligence/



Half of AI researchers believe there is a >10% chance of human extinction due to uncontrolled AGI

The AI Dilemma
https://vimeo.com/809258916/92b420d98



https://arxiv.org/abs/2304.11082

Today's alignment methods are too "soft"

[Submitted on 19 Apr 2023]

Fundamental Limitations of Alignment in Large Language Models

Yotam Wolf, Noam Wies, Yoav Levine, Amnon Shashua

in large language models. Importantly, we prove that for any behavior that has a finite probability of being exhibited by the model, there exist prompts that can trigger the model into outputting this behavior, with probability that increases with the length of the prompt. This implies that any alignment process that attenuates undesired behavior but does not remove it altogether, is not safe against adversarial prompting attacks. Furthermore, our framework hints at the mechanism by which leading alignment approaches such as reinforcement learning from human feedback increase the LLM's proneness to being prompted into the undesired behaviors.

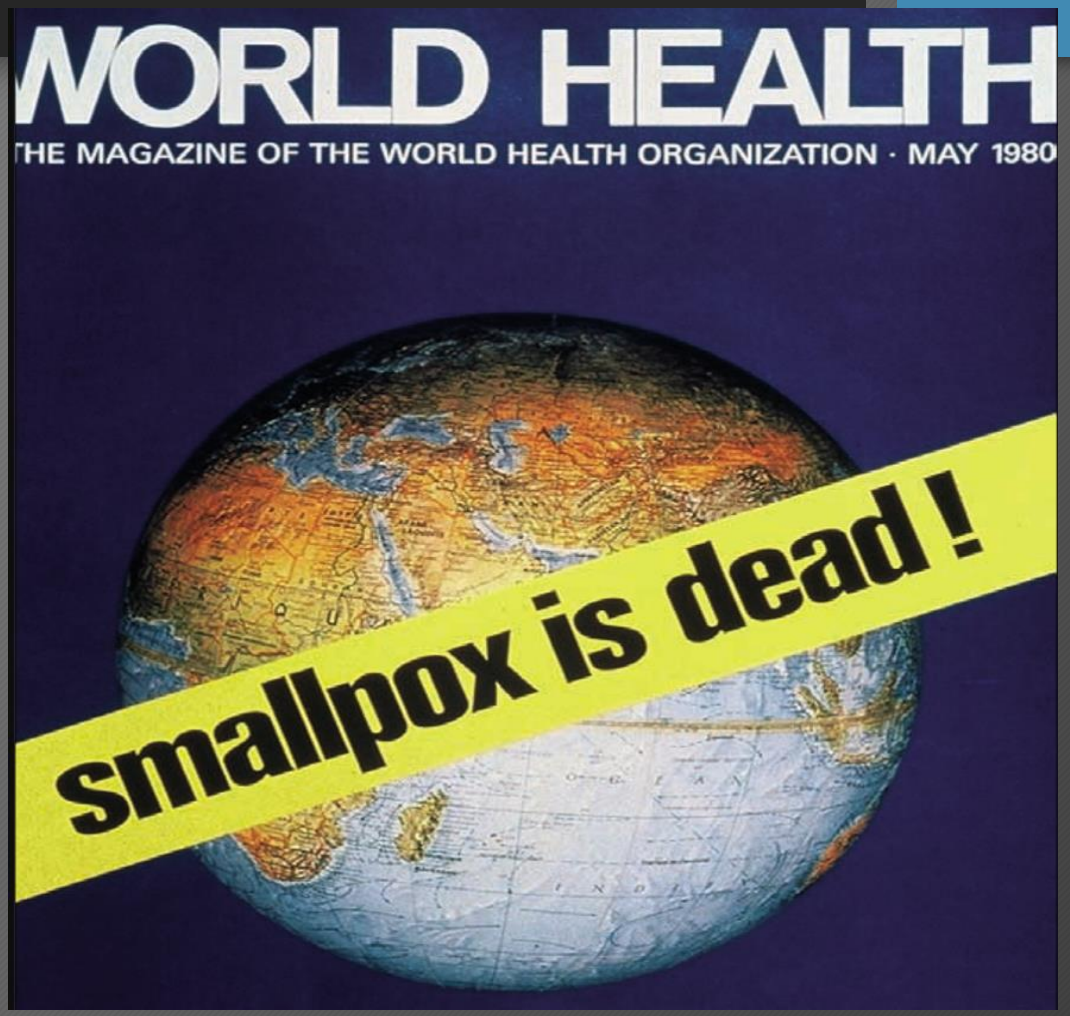
We need adversarial guarantees, not just probabilistic!

We need absolute guarantees that AGI won't:

Nuke-launching AI would be illegal under proposed US law

Markey and Lieu seek to ban fed funds for nuke launches without "meaningful human control."

BENJ EDWARDS - 4/28/2023, 9:11 AM



https://arstechnica.com/information-technology/2023/04/nuke-launching-ai-would-be-illegal-under-proposed-us-law/

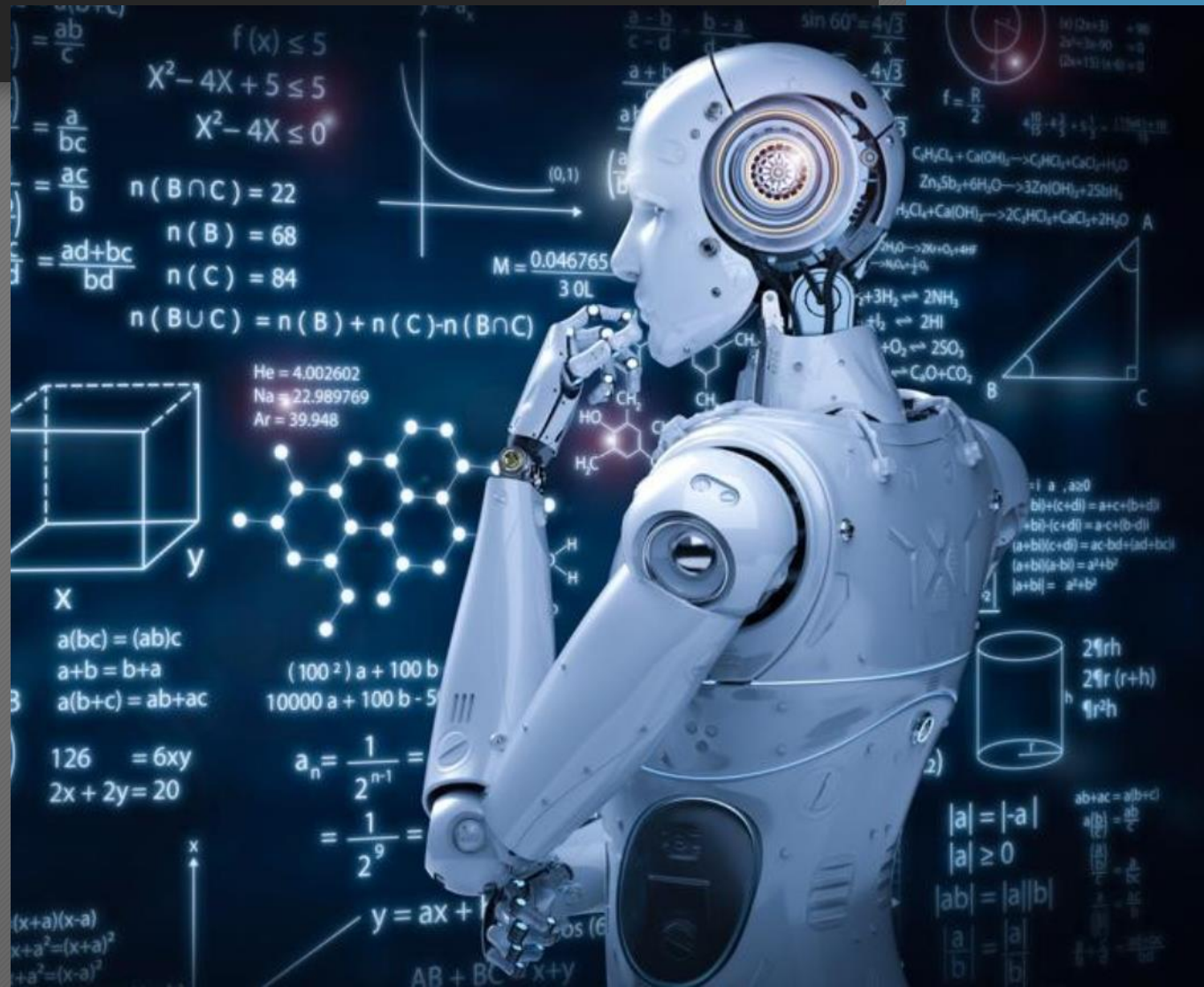
Muehlhauser (OpenPhil) AI Policy Proposals:

- Monitoring and remote shutdown of cutting-edge AI chips
- Rapid shutdown of large compute clusters and training runs
- Tracking and licensing of cutting-edge AI chips
- Export control of AI trained with $1 billion compute
- License large cluster formation
- License frontier AI model development
- Infosec on frontier AI models to prevent proliferation
- AI incident reporting

https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/

Who flips the switch to turn off an AGI?

- Humans?
  - Corruptible
  - May not fully understand criteria
  - Too slow!
- AGIs?
  - Risk of deception
  - How can we trust it?
- Crypto “smart contracts”?
  - Closer!
  - Need guarantees
- Proven “formal contracts”!



Humanity's most powerful safety technology: Mathematical Proof

- 350BC Origins: Aristotle, Euclid
- 1637 Mathematical Analysis: Descartes, Weierstrass
- 1854 Modern Logic: Boole, Cantor, Frege
- 1925 Set Theory: Zermelo, Fraenkel
- 1934 Type Theory: Curry, Godel, Barendregt
- 1936 Computation: Turing, Church
- 1975 Standard Model of Physics
- We now have formal models for: All of Mathematics, Physics, Computer Science, Engineering, Economics, …
- Proof provides absolute guarantees within a formal model!

Automated Theorem Provers

- 1956 Propositional Theorem Provers: Newell
- 1976 First Order Theorem Provers: Luckham
- 2000 SAT/SMT Solvers: SAT Competitions
- 2000 Proof Assistants: HOL, Mizar, MetaMath, Coq, Lean, Isabelle, …
- 2020 Neural Theorem Provers: GPT-F, Hypertree HTPS

http://us.metamath.org/mpeuni/mmset.html

E.g. MetaMath

- Simple ZFC foundations sufficient for formalizing everything!
- Fast 300 line Python proof checker
- “de Bruijn factor”: Formal statements are ~4X size of English statements
- 38K proven theorems
- Basic set theory, real and complex analysis, number theory, category theory, abstract algebra, linear algebra, topology, geometry, graph theory, Hilbert spaces, etc.

Axioms of propositional calculus

Axiom *Simp*	ax-1	$\vdash (\varphi \rightarrow (\psi \rightarrow \varphi))$
Axiom *Frege*	ax-2	$\vdash ((\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi \rightarrow \psi) \rightarrow (\varphi \rightarrow \chi)))$
Axiom *Transp*	ax-3	$\vdash ((\neg \varphi \rightarrow \neg \psi) \rightarrow (\psi \rightarrow \varphi))$
Rule of Modus Ponens	ax-mp	$\vdash \varphi \;\; \& \;\; \vdash (\varphi \rightarrow \psi) \;\; \Rightarrow \;\; \vdash \psi$

Axioms of predicate calculus with equality - Tarski's S2

Rule of Generalization	ax-gen	$\vdash \varphi \;\; \Rightarrow \;\; \vdash \forall x \varphi$
Quantified Implication	ax-4	$\vdash (\forall x(\varphi \rightarrow \psi) \rightarrow (\forall x \varphi \rightarrow \forall x \psi))$
Distinctness	ax-5	$\vdash (\varphi \rightarrow \forall x \varphi)$, where x does not occur in φ
Existence	ax-6	$\vdash \neg \forall x \neg x = y$
Equality	ax-7	$\vdash (x = y \rightarrow (x = z \rightarrow y = z))$
Left Equality for Binary Predicate	ax-8	$\vdash (x = y \rightarrow (x \in z \rightarrow y \in z))$
Right Equality for Binary Predicate	ax-9	$\vdash (x = y \rightarrow (z \in x \rightarrow z \in y))$

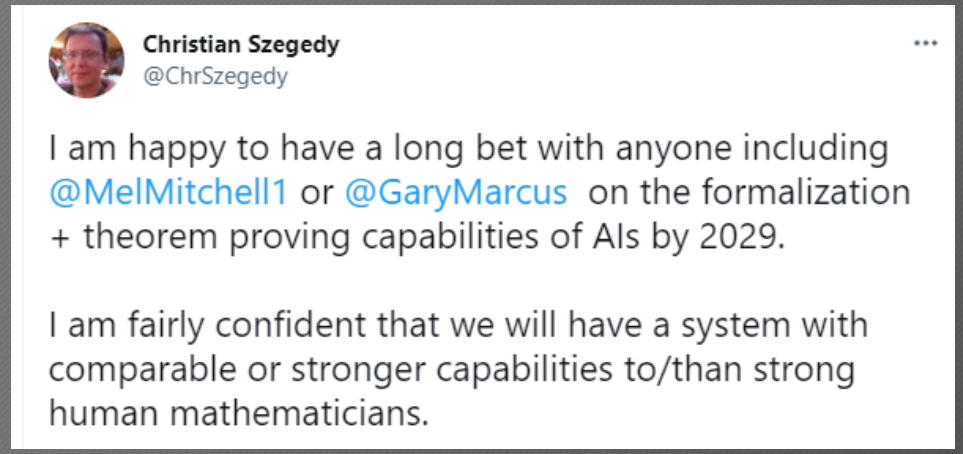
Zermelo-Fraenkel Set Theory with Choice (ZFC)

Axiom of Extensionality	ax-ext	$\vdash (\forall z(z \in x \leftrightarrow z \in y) \rightarrow x = y)$
Axiom of Replacement	ax-rep	$\vdash (\forall w \exists y \forall z(\forall y \varphi \rightarrow z = y) \rightarrow \exists y \forall z(z \in y \leftrightarrow \exists w(w \in x \wedge \forall y \varphi)))$
Axiom of Power Sets	ax-pow	$\vdash \exists y \forall z(\forall w(w \in z \rightarrow w \in x) \rightarrow z \in y)$
Axiom of Union	ax-un	$\vdash \exists y \forall z(\exists w(z \in w \wedge w \in x) \rightarrow z \in y)$
Axiom of Regularity (Foundation)	ax-reg	$\vdash (\exists y\, y \in x \rightarrow \exists y(y \in x \wedge \forall z(z \in y \rightarrow \neg z \in x)))$
Axiom of Infinity	ax-inf	$\vdash \exists y(x \in y \wedge \forall z(z \in y \rightarrow \exists w(z \in w \wedge w \in y)))$
Axiom of Choice	ax-ac	$\vdash \exists y \forall z \forall w((z \in w \wedge w \in x) \rightarrow \exists v \forall u(\exists t((u \in w \wedge w \in t) \wedge (u \in t \wedge t \in y)) \leftrightarrow u = v))$

https://www.sciencedirect.com/science/article/pii/S1570868305000686

Transformers for Theorem Proving and Autoformalization

- 2020 GPT-f : Trained on 36K MetaMath theorems and 3M proofsteps, proves 56.5% of held out theorems
- 2022 HyperTree Proof Search : AlphaZero-style MCTS transformer, trained on 40GB of arXiv math, proves 82.6% of held out MetaMath theorems
- 2022 Autoformalization with Large Language Models



https://twitter.com/ChrSzegedy/status/1534082344096702464

Distilling Formal Models from Deep Nets

Abstract interpretation, Interval Methods, Taylor Series bounds,…
Generalizations of BackProp to intervals, relations, constraints,…

- Introduction to Neural Network Verification
  https://arxiv.org/abs/2109.10317
- Neuro-Symbolic Verification of Deep Neural Networks
  https://arxiv.org/abs/2203.00938
- AutoBound: Automatically Bounding the Taylor Remainder Series: Tighter Bounds and New Applications
  https://arxiv.org/abs/2212.11429

“Proof for Safety” Insights

- Proof checkers can be tiny, fast, and absolutely reliable.
- Small, interpretable, proven systems can control powerful AIs.
- Simple ZFC or type theory can encode all real-world systems.
- Undecidability (e.g. halting) doesn’t matter – only use proven systems.
- Formal ontology is critical – layered abstractions based in physics.
- Incorporate blackboxes by forcing them to generate proofs for actions.
- Never let AGI or humans directly act on dangerous systems.
- Use AGIs to generate whitebox systems provably obeying social contracts.
- Society is guardrailed by a network of proven contracts.

“Proof for Safety” Challenges

- Abstract levels must be implemented at lower levels (e.g. Rowhammer).
- Need formal models of existing systems.
- Need “Formal Guardrails,” challenging but simpler than full alignment rules.
- Critical hardware must have full provable provenance to avoid hardware Trojans.
- Critical hardware must use provably unbreakable cryptography (e.g. OTP).
- Social acceptance of “Provably Secure Sensing” may be challenging.
- Transitioning from today’s social systems to provably secure ones.
- Political process of transitioning to proven systems.

Vision for a "Proof-Protected Society"

- Every human has a "Personal AI" which provably represents their interests.
- "Provably Secure Sensing" detects risky acts without leaking private info.
- "Semantic voting" provably aggregates individual needs at the societal level.
- A "Formal Constitution" provably governs all contracts.
- All contracts are formal, automatically efficiently negotiated and enforced.
- "Precise guardrails" guarantee safety with full freedom beyond that.
- Leads to human flourishing in an AGI world of abundance.