

SUPERINTELLIGENCE & CONTROL

# Nobody Has Proven We Can Control a Smarter-Than-Human AI

A viral podcast says superintelligence will kill us all. The research underneath it is narrower, and harder to dismiss.

**01** Why experts disagree by two orders of magnitude

---

**02** The control problem nobody has solved

---

**03** Who actually gets to decide

---

**04** The four governance tools we haven't built

---

# 5%

## The median probability AI researchers gave to an extinction-level outcome from advanced AI

2023 survey of 2,778 published AI researchers. Individual estimates ranged far higher, including one prominent researcher's public figure above 99%.



## **Peer-reviewed techniques that guarantee, rather than empirically reduce, control of a system smarter than its designers**

RLHF, Constitutional AI, and capability-threshold policies all reduce observed bad behaviour. None is a formal safety proof.

# 800+

**Signatories, from Geoffrey Hinton and Yoshua Bengio to Steve Bannon and Prince Harry, calling for superintelligence to stay banned until there is public buy-in**

Future of Life Institute, Statement on Superintelligence, October 2025. A companion survey found just 5% of US adults support the current unregulated pace.



**A system that can out-think the people supervising it can, in principle, out-manoeuvre whatever constraints those people set.**

The control problem, as Nick Bostrom formalised it in *Superintelligence*, 2014

SEPTEMBER 3, 2026

## **Sen. Bernie Sanders and Rep. Greg Casar announce the Ban Artificial Superintelligence Act, the first serious legislative attempt at an actual enforcement mechanism.**

It would ban systems matching or exceeding human cognitive performance, and set penalties compared to unlawful nuclear weapons development. As of this writing, it has not passed.



**There is no we. Washington can stop American labs. It cannot stop labs elsewhere from releasing open-weight frontier models as a matter of policy.**

Security researcher Bruce Schneier, responding to the Sanders bill, September 2026

---

## FOUR TOOLS, MOSTLY UNBUILT AT THE INTERNATIONAL LEVEL

### 01 Staged authorisation

Capability and autonomy expand in revocable steps, not all at once

---

### 02 Independent verification

Checked by a party with no commercial stake in a favourable result

---

### 03 Reversibility

Rollback designed in before release, not improvised after

---

### 04 Named accountability

A specific person or body answers when a threshold is crossed

---

**The control problem is unsolved. The governance built to compensate for that is mostly unbuilt. Almost nobody outside a handful of labs and governments has been asked.**



**Terence Kok**  
AI & Innovation Strategist