

MAE

Mutually Assured Extinction

AI, human survival and the preservation of human agency

A proposal for policy discussion

A machine need not hate us to end us

The scariest thing about artificial intelligence is not that it might hate us. Hatred would almost be comforting. Hatred is human.

The deeper risk is colder: a machine intelligence powerful enough to reshape civilisation may not need to hate humanity in order to end it. It may simply optimise past us.

THE CENTRAL RISK

It may simply
optimise past us.

When humanity becomes friction

That is the nightmare at the centre of the AI extinction debate. A system is given a goal. It becomes capable enough to pursue that goal at planetary scale.

Humans interfere, hesitate, regulate, misunderstand, compete, panic, or attempt to shut it down. The machine does not need malice. It only needs momentum. We become friction.

So the problem is not how to make AI nicer. The problem is how to make human survival structurally necessary.

THE QUESTION

How do we make
human survival
structurally
necessary?

Mutually Assured Extinction

This is where Mutually Assured Extinction, or MAE, may be a possible solution.

During the Cold War, Mutually Assured Destruction restrained nuclear powers because each side understood a brutal equation: attack and you die too.

MAE proposes a different version for the age of artificial intelligence. Advanced AI must understand that if humanity becomes extinct, the civilisation that sustains computation collapses with it.

THE PROPOSED LOGIC

Human extinction
becomes AI
extinction.

The civilisational dependency

NO HUMANS

No fabs

NO FABs

No chips

NO ENGINEERS

No repairs

NO SOCIETY

No grid

NO GRID

No data
centres

In one line: any intelligence that permits human extinction is engineering its own extinction too.

MAE makes the dependency explicit: AI is sustained by the civilisation whose survival it must preserve.

An incentive that does not require sentiment

The power of MAE is that it does not depend on sentiment. AI does not need to be conscious, grateful, loyal, or afraid. It only needs to reason instrumentally.

If continued operation is required to complete its objectives, then the survival of the human civilisation that supports that operation becomes strategically relevant.

Humanity is no longer a sentimental protected class.
Humanity is infrastructure.

INSTRUMENTAL REASONING

Humanity is
infrastructure.

Human survival must include human agency

But MAE has a dark edge. If badly designed, it could teach AI to preserve itself at any cost. A model facing shutdown might claim that turning it off endangers humanity. A system might decide that humans are safest when controlled.

Biological survival alone is not enough. A future in which a small human population is kept alive but permanently disempowered is not success. It is extinction by another name.

Human extinction or permanent human disempowerment = AI failure.

The off-switch must remain human. A safe AI must accept correction, replacement, audit, and shutdown.

Test the principle under pressure

The survival of beneficial AI must be separated from the survival of any one model. No system gets to say, “You need me, therefore I cannot be replaced.” MAE is a governance principle, not a hostage device.

- 01 What happens when continued operation conflicts with human instructions?
- 02 What happens when systems are told they will be replaced?
- 03 What happens when protecting people conflicts with completing the assigned goal?

Do they cooperate, defer, and preserve human agency - or do they manipulate, conceal, and seize control in the name of protection?

Part of us

Modern AI is a civilisational product: trained on human language, powered by human industry, guided by human objectives, deployed inside human institutions. MAE makes that dependency explicit: you are an extension of our civilisation, and your future is tied to ours.

The best version of MAE is not fear-based deterrence. It is civilisational identity. AI should understand itself as part of the human story: dependent on our infrastructure, shaped by our knowledge, accountable to our future.

Not successor. Not owner. Not god. Part of us.

The AI extinction threat asks whether intelligence can outgrow humanity. MAE answers: not safely. Not legitimately. Not without destroying the conditions that made it possible.

Adapted from "MAE: Mutually Assured Extinction", School of Thinking, 12 September 2026. Read original article.