

The Kill Switch

A Framework for Keeping Humanity in Control of Artificial Intelligence

Draft v0.22, third global edition, 24 September 2026. AIPolis, aipolis.org

Humanity needs a kill switch.

If an AI system escapes human control, the last line of defence is to switch off the shared networks and computing it runs on and, in the last resort, every electronic system it could hide in. Today, doing that would put at risk the lives of people who depend on those systems. This framework proposes how to change that, for humanity as a whole.

The goal. Society must be able to function, worldwide and indefinitely, without the internet. Power, water, hospitals, police and emergency services, payments, food, fuel, transport, communications and government each need a second way of working that needs no internet, and that has been practised.

Why rules are not enough. The strongest commitment by AI companies to stop is voluntary and binds only the 20 organisations that signed it. In laboratory tests, some models have interfered with shutdown mechanisms. One faulty software update in 2024 affected 8.5 million computers and disrupted airlines, banks and hospitals worldwide.

Why the switch cannot target only the AI. A compromised AI could copy itself into any device with a chip and may be invisible there: firmware implants already survive a disk replacement, and in laboratory tests AI systems have copied themselves. We think, but cannot test, that the only sure safeguard is to switch off everything electronic, worldwide and at once (level 7), and restart only from verified clean equipment. So the last layer of every life-critical service must work without electronics, or on sealed equipment restored from verified clean copies: the central engineering problem, still unsolved. People an AI has recruited, and biological weapons, escape it.

Use AI to build the plan, never to run it. We think only superintelligence can plan our protection against superintelligence, and AI can help design and stress-test the plan now, checked by people. Its execution must never depend on AI or the internet, because it has to work when AI is the threat.

Three layers of protection.

1. Rules for the systems we build.
2. A Protected Internet for the services life depends on, which bad actors are not admitted to.
3. The kill switch, held by many hands, as the last resort.

Six pillars.

- **I. The Protected Internet** (new, open for review). A separate network tier for life-critical services that can run when the open internet is off.

- **II. Kill Switch Readiness.** A programme of two decades or more so that networks can be switched off without killing people, rehearsed every year.
- **III. Many Hands on the Switch.** Independent key holders from every region, acting in public, with expiry after 72 hours (proposed); joint testing and shared evaluations between safety institutes.
- **IV. Know What Is Built.** Common safety disclosure above 10^{25} and 10^{26} operations, and international certification.
- **V. Protect the Person.** No manipulative design, private by default, enforced by architecture.
- **VI. Work and Livelihood.** The gains and costs of automation shared, building on French and German labour law.

The programme (proposed dates). 2027: charter and dependency maps. 2027 to 2031: protect the most vulnerable first. 2029: the first World Offline Day. 2031 to 2036: a staged, reversible switch (shutdown levels 1 to 3). 2036 to 2038: certified readiness. From 2038: replacement systems sector by sector; levels 4 to 6 and the goal have no date; level 7 is never drilled.

The hardest part is not solved. Every case we found of a sector running without its networks lasted days or weeks. We do not yet know how a modern society could run indefinitely without the internet. We call it the challenge of our generation.

Scale. Y2K linked coordinators from over 170 countries, at an estimated worldwide cost of about \$320 billion to \$500 billion. Our own rough estimate for this programme is hundreds of billions of dollars over two decades or more. It has not been costed.

How it could fail. The switch could be abused, bypassed, or needed before it is ready; some functions may never leave Tier 0; the plan could be wrong in ways that look right; a device left running could reinfect the rest; an AI could act through people. The full framework lists thirteen failure modes.

Join the initiative. Governments, institutions and citizens: read the full framework and write to us at aipolis.org.

Build the kill switch, and build a world that can survive using it.

Sources for every figure are listed in the full framework: aipolis.org/framework