

The Kill Switch

A Framework for Keeping Humanity in Control of Artificial Intelligence

Draft v0.23, fourth global edition, 28 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, or what it built, could live in. Today, doing that would put at risk the lives of people who depend on those systems. This framework proposes how to change that.

The goal. Society must be able to function, worldwide and indefinitely, without the internet. Every life-critical service needs 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.

Why the switch cannot target only the AI. An AI that expects to be switched off could first build destructive things that are not AI: malware, scheduled automations, triggers set to fire if it goes silent. Switching off the AI does not stop them. People have planted timed code, and criminals have used AI to write ransomware; we found no AI doing so on its own. So the switch must take away the AI's oxygen: everywhere it, or what it built, could run or hide. We think, but cannot test, that only a worldwide switch-off of everything that can be infected (level 7), then a verified clean restart, could be complete. Released biological weapons, and people an AI has recruited, escape it. The goal is that no one is harmed: uninfected equipment (no data connection, no rewritable code; not tamper-proof) keeps running, with a public gap list whose target is zero.

Use AI to build the plan, never to run it. We think only superintelligence can plan our protection against superintelligence: AI can help design and stress-test the plan now, checked by people, but its execution must never depend on AI or the internet: it must 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, under a decision rule that weighs harm on both sides; every use expires unless renewed (proposed).
- **IV. Know What Is Built.** Common safety disclosure above 10^{25} and 10^{26} operations, international certification, and monitoring inside each lab.
- **V. Protect the Person.** No manipulative design, private by default, enforced by architecture.
- **VI. Work and Livelihood.** The gains and costs of automation shared.

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 sector we found running without its networks did so for 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. This programme has not been costed. Our first rough range is hundreds of billions to trillions of dollars over two decades or more, depending mostly on how many households get their own backup power. Those who did not build frontier AI should not pay to survive it, and the people most at risk are protected first (Annex E).

How it could fail. The switch could be abused, falsely triggered, bypassed, or needed before it is ready; anything left running could bring it back; an AI could act through people. The full framework lists twenty-one failure modes.

Join the initiative. Governments, institutions and citizens: 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