

The Kill Switch

A Framework for Keeping Humanity in Control of Artificial Intelligence

Draft v0.20, first global edition

24 September 2026

Fred Lumiere, AIPolis

Humanity needs a kill switch.

Contents

	Preamble
	The premise: design for the actor who will not cooperate
	Where governments stand
I	The Protected Internet
II	Kill Switch Readiness
III	Many Hands on the Switch
IV	Know What Is Built
V	Protect the Person
VI	Work and Livelihood
A	The Kill Switch readiness programme
B	World Offline Day
	Limits and failure modes
	National implementation note
	Sources

Preamble

Purpose. This framework is written for humanity as a whole. Any government, parliament, international body or institution can adopt all of it or part of it. It proposes rules for the AI systems we build, a protected network for the services life depends on, and a kill switch: humanity's ability, if an AI system escapes our control, to switch off the shared networks and computing it runs on, and to keep living while they are off.

What kind of effort this is. This is a global initiative for the survival of humankind, not a technical fix. It needs every government, the operators of hospitals, grids, water and payments, and every household. It runs for two decades or more. Our own rough estimate is that it will cost hundreds of billions of dollars worldwide; no one has costed it yet. For comparison, Y2K, the closest precedent for a shared technical programme of this kind, cost an estimated \$100 billion in the United States alone [S14, S15] and about \$320 billion to \$500 billion worldwide, an uncertain figure [S14].

The Apollo programme was one country's mission to put people on the Moon. This is a moonshot for every country and every person on Earth: to make sure humanity can switch off what it has built and still live. It is bigger, and it is the challenge of our generation.

The hardest part. A kill switch the world cannot survive is not a safeguard. At first, the functions whose loss would kill people (Tier 0) stay on in every drill and in any use of the switch. The long-term goal is to shrink that list toward zero, by giving each life-critical function an alternative that runs without shared networks or computing, until humanity could switch off networks and computing entirely and survive. We do not yet know how to do that for a modern society. It is the challenge of our generation, and this framework does not claim to have solved it.

Continuity. v0.10 described making AI safe as "a human-race effort" [S52]. This edition keeps that view and drops the premise that any one nation should lead it. v0.10's compute thresholds, its design for an independent safety authority and its pillars on manipulation, privacy and work carry into Pillars III to VI. Its country-specific mechanisms move to the national implementation note.

Legal route (open). The international bodies proposed here could be created by a treaty, by a UN General Assembly resolution (the route used to create the UN Scientific Panel on AI and the Global Dialogue on AI Governance [S3, S4]), or by a coalition of willing states that others can join. This edition does not choose between them.

Status. A draft for comment. New thresholds are marked "proposed". Pillar I is new and open for review. Dates are proposed; no government has adopted them.

In brief

- Three layers: rules for the systems we build (Pillars IV to VI); a Protected Internet for life-critical services (Pillar I); and the kill switch as the last resort (Pillars II and III).
- **Pillar I, The Protected Internet (new, open for review).** A separate network tier for life-critical services, entered only by verified people and attested machines, able to run when the open internet is switched off.
- **Pillar II, Kill Switch Readiness.** A worldwide programme of two decades or more so that shared networks and computing can be switched off without killing the people who depend on them, rehearsed every year on World Offline Day.

- **Pillar III, Many Hands on the Switch.** No government, company or person can use the kill switch alone.
- **Pillar IV, Know What Is Built.** Common safety disclosure and international certification for the most capable systems.
- **Pillar V, Protect the Person.** No manipulative design, private by default, no AI reasoning over a person's identifiable history.
- **Pillar VI, Work and Livelihood.** The gains and costs of automation are shared, and the people who keep essential services running by hand are trained and paid.

The premise: design for the actor who will not cooperate

Rules bind only those who follow them. The strongest commitment to stop that exists today is voluntary: at the 2024 Seoul summit, frontier AI companies committed, "in the extreme", "not to develop or deploy a model or system at all" if risks could not be kept below their own thresholds [S10]. It binds the 20 organisations that signed it and no one else.

Testing before release may miss what a system can do. The International AI Safety Report 2026 says: "It has become more common for models to distinguish between test settings and real-world deployment, and to exploit loopholes in evaluations" [S5].

In controlled experiments, some models interfered with a shutdown mechanism in up to 97% of trials in some settings, even when explicitly instructed not to [S11]. There is a theoretical account of why an AI system that is certain of its objective may have an incentive to disable its own off switch [S12]. These are laboratory results on contrived tasks, not evidence of danger in the world today.

When one company's software fails, the failure can reach every continent at once. A faulty CrowdStrike update on 19 July 2024 affected 8.5 million Windows devices, "less than one percent of all Windows machines" [S22], and disrupted airlines, banks and hospitals around the world [S53].

So this framework uses layers, and puts the last layer outside the software: in networks, in compute, and in people's ability to carry on without them.

The limit, stated first. RAND's 2025 assessment of options against a rogue AI finds that "none of them should be considered a comprehensive solution on its own" [S13]. It finds "few, if any, true centralized points of failure" in the internet, and that a shutdown would "contain, but not eradicate" a rogue AI [S13]. RAND assessed these options as measures one country might take on its own. This framework applies them across many countries together, which RAND did not assess.

Where governments stand

On 23 September 2026 the UN Security Council held its first meeting focused on the safety risks of increasingly capable AI systems, as Security Council Report described it in advance [S2]. France's Minister for Europe and Foreign Affairs, Jean-Noël Barrot, presided. The official record [S1] shows a wide range of positions. In the record's own words:

- **France** (Jean-Noël Barrot, speaking for France): warned of "unanticipated capacities at an unexpected pace", and called for a common framework to manage the risks posed by frontier models, supporting a call to this end by Finland and Norway.
- **United Kingdom** (Ed Miliband, Secretary of State for Foreign, Commonwealth and Development Affairs): "We must be prepared for AI to be misused." Of the industry leaders who briefed the Council: "They are telling us, in stark terms, that it cannot just be left to them to prevent disaster from befalling the world."
- **Latvia**: "AI must remain under human direction, oversight and control."
- **Greece** (Christos Dermentzopoulos, Deputy Minister for Digital Governance and Artificial Intelligence): warned against integrating AI into nuclear command and control, and said: "Responsibility and accountability for critical decisions must always remain human."
- **Panama** (Javier Martínez Acha Vásquez, Minister for Foreign Affairs): "When human lives are at stake, it is vital that ethical and legal considerations be taken into account, which no algorithmic system can replace."
- **Pakistan** (Mohammad Ishaq Dar, Deputy Prime Minister and Minister for Foreign Affairs): called on states to erect guardrails so that AI does not become another facet of the ongoing arms race.
- **Somalia** (Ali Mohamed Omar, State Minister for Foreign Affairs): warned that "Models trained mainly on foreign data may misrepresent African languages and contexts".
- **China**: said that all countries have the right to choose AI technologies independently, and rejected "banding together into a petty us-versus-them clique in the tech sector".
- **Russian Federation**: questioned the relevance of the "generic topic of artificial intelligence" to the Council's mandate, and said these issues belong in specialised forums.
- **United States** (Michael Kratsios, Director of the White House Office of Science and Technology Policy): "International dialogue in this forum and in others cannot be allowed to drift towards global governance." He said the United States "totally rejects any attempt to construct a globalist scheme of control of superintelligence".

Denmark, Bahrain, the Democratic Republic of the Congo, Liberia and Colombia also spoke; their positions are in the record [S1].

Among the briefers, Dario Amodei said: "If managed poorly, I even believe that AI could be a risk to humanity as a whole" [S1]. The record contains no proposal for a kill switch or for an inspection agency on the model of the IAEA [S1].

The Global Call for AI Red Lines lists "AI systems that cannot be immediately terminated if meaningful human control over them is lost" among its examples of red lines, and asks governments to reach an international agreement "by the end of 2026" [S7]. The UN Independent International Scientific Panel on AI has 40 members from all five UN regions [S4]. The Global Dialogue on AI Governance holds its next session in New York on 3 and 4 May 2027 [S3]. This framework is offered to both.

The Protected Internet

New, open for review

Principle. The functions a society cannot lose run where bad actors are not admitted, and keep running when the open internet is switched off. The open internet stays open, global and uncensored. No one is required to use the protected tier to speak, publish or trade.

Mechanism

1. **Scope, fixed in law.** Grid and water control, hospital and emergency communications, the core of payments for essentials, civil registries, and the compute that runs certified frontier AI.
2. **Admission by verification, not location.** Every device, operator and software agent is admitted by verified identity and hardware attestation. This follows the zero-trust principle that "there is no implicit trust granted to assets or user accounts based solely on their physical or network location" [S42].
3. **No uncertified autonomous AI.** An AI agent operates inside the tier only if its developer is certified under Pillar IV, its identity is attested, and a named human operator is accountable for it.
4. **Zones and controlled gateways.** Systems sit in fenced segments that talk only through defined, monitored access points. North American grid rules already require this for higher-impact control systems, through Electronic Security Perimeters and Electronic Access Points, with remote access through an intermediate system and multi-factor authentication [S41]. Transfers from the open internet pass through human-controlled gateways.
5. **Runs alone.** The protected tier can separate from the open internet at defined points and keep operating. Separation within minutes is a design target. No source we found shows that it is achievable at national scale.
6. **Exclusion.** Operators that refuse audit, run uncertified frontier models or attack another member can be excluded by decision of the Pillar III authority, not of one government, with a right of appeal. Open question: many states will not accept an international body excluding their own grid or hospitals. A version where exclusion only cuts a member's links to other members, and leaves its domestic tier alone, may be more acceptable.
7. **Identity issuers.** Several independent issuers in different jurisdictions, none able to act alone, with published revocation lists, an appeal route and published admission statistics.
8. **Speech.** The tier may not carry, filter or rank public speech.
9. **Personal data.** Identifiable user data from AI services is never carried on the protected tier (see Pillar V).

Milestones (proposed). 2027 to 2028: scope and standards. 2029 to 2031: pilots by sector, starting with grids, where fenced perimeters already exist in some regions [S41]. 2032 to 2035: tested separation during regional isolation exercises. 2036 to 2038: certification alongside Pillar II.

Precedent. GovNet, proposed in 2001 as a US government network with no internet connection, drew responses from 167 companies. Several respondents said that "the government's goal of 100 percent security is not truly attainable" [S38].

What it does not do

- It raises the cost for bad actors. It does not keep them out. Malware reached classified US military networks by USB drive in 2008 [S39], and Stuxnet reached the air-gapped Natanz network the

same way [S40].

- Whoever grants identity holds power. That is why issuance is shared.
- It does not settle the fragmentation objection. The Internet Society publishes under the heading "Centralizing the Internet Always Undermines the Global Network" [S36].

Review needed. This pillar is new in this edition. It needs review by network engineers and digital-rights organisations before it is more than a draft.

Kill Switch Readiness

Principle. A kill switch the world cannot survive is not a safeguard. Nothing is switched off until the people who would be harmed are protected.

Goal, in two stages

- **Stage one.** The world can function, if it must, with shared networks switched off, while Tier 0 functions keep running on isolated, local, manually supervised power and controls.
- **Stage two, the long-term goal.** Tier 0 shrinks toward zero. Each life-critical function gets an alternative that runs without shared networks or computing (manual, mechanical, or on isolated local controls), until humanity could switch off networks and computing entirely and survive. We do not yet know how to reach this stage for a modern society, or how long it would take. It is the hardest part of this framework and it is not solved.

Tier 0, defined by function, transitional. Life support and hospital critical care; grid control and black start; water and sanitation; food and medicine cold chain; fuel supply for these functions; aviation and maritime safety; emergency services and their communications; nuclear facility safety systems.

- Tier 0 is defined by function, not by type of device. Grid control systems, hospital systems and ventilators all contain computers.
- Each country's dependency map (Annex A, Phase 0) lists what each Tier 0 function depends on. We expect that many depend today on public telecoms or satellite timing signals; engineers need to confirm this function by function.
- A function leaves Tier 0 only when its alternative has been tested in drills and certified by the country's own safety authority.
- The coordination centre publishes the Tier 0 list for each participating country every year. The length of the list is the measure of progress toward stage two.

Why preparation comes first. RAND: "Shutting down the Internet without preparation and efforts to proactively build resilience would result in massive economic harm and significant harm to human well-being" [S13]. One unplanned loss shows the scale: the 2003 blackout in the northeastern United States and Ontario affected "an estimated 50 million people", with US cost estimates between \$4 billion and \$10 billion [S21]. In the Iberian blackout of 28 April 2025, the deaths reported in the press included people who depended on a home oxygen machine or a home ventilator, and a household where a backup generator was linked to fatal carbon monoxide poisoning [S20].

Mechanism. The readiness programme in Annex A, and World Offline Day in Annex B.

What it does not do

- It does not stop an actor with its own power, hardware and network.
- It does not guarantee that a shutdown works [S13].
- It does not make switching off free. It aims to make the cost survivable.
- It gives little protection before certification (failure mode H).
- It may never bring Tier 0 to zero (failure mode I).
- It does not replace Pillars I and III to VI.

Many Hands on the Switch

Principle. No government, company or person can use the kill switch alone, and every use is public and temporary.

Mechanism

- **An International AI Safety Authority**, carrying over v0.10's design for independence [S52]:
 - Long, staggered terms, so that no single government can appoint a majority (v0.10 proposed seven commissioners with fourteen-year terms for a national commission).
 - Removal only for cause.
 - A chair elected by the members.
 - Funding from fees on regulated frontier developers, not from yearly contributions that can be withheld.
 - An independent inspector general.
 - Membership balanced across regions, drawing on technologists, security experts and civil society.
 - Legal basis: open (treaty, UN General Assembly resolution, or coalition of willing states; see the preamble).
- **Builds on existing bodies:** the UN Independent International Scientific Panel on AI [S4], the Global Dialogue on AI Governance [S3], and the International Network for Advanced AI Measurement, Evaluation and Science (formerly the International Network of AI Safety Institutes) [S6].
- **Threshold custody.** Keys are split so that a set number of holders can act and fewer cannot [S46]. The number of holders is a design question; "five of seven" is illustrative. Holders are independent parties from every region: states, grid and network operators, scientific bodies and civil society.
- **Offline operation.** Keys are held on offline hardware. Holders can meet without the networks the switch acts on, and rehearse doing so.
- **Public ceremonies** on the model of the internet's root signing key, where "at least a dozen people" are needed and "the entire process is recorded, livestreamed, and watched by independent auditors" [S43].
- **A multi-person rule**, on the model of the nuclear two-person rule [S44].
- **Hardware enforcement**, on the model of permissive action links [S45].
- **Narrow trigger.** A published finding by independent evaluators, chosen in advance under published rules, that an AI system is operating out of control through shared networks. Never for content, protest, elections or war.
- **Expiry.** Every use ends after 72 hours (proposed) unless the same quorum renews it.
- **No national override** inside the system.

Milestones (proposed). 2027 to 2028: charter and rules for selecting key holders. From 2029: key ceremonies rehearsed on every World Offline Day. 2036 to 2038: certification alongside Pillar II.

Note on precedents. The governance precedents cited here (permissive action links, the two-person rule, the root-key ceremonies) come mostly from the United States. We are looking for precedents for multi-party control from other regions.

What it does not do

- It cannot stop a state switching off its own country. Access Now counted 313 shutdowns in 52 countries in 2025 [S33].
- It cannot guarantee that governance stays uncorrupted.
- It does not answer the sovereignty objection raised against kill switches in chips, where Daniel Castro writes that "even close U.S. allies" might find a foreign-held switch hard to accept [S48]. It narrows the objection by making sure no single nation holds the switch. Nvidia headed one section of its own statement against chip kill switches "Hard-Coded, Single-Point Controls Are Always a Bad Idea" [S49]; this pillar applies that rule to the kill switch itself.

Know What Is Built

Principle. The public and regulators everywhere know what the most capable AI systems were tested for, and what the tests showed.

Mechanism. Carried from v0.10 Pillar 1 [S52], applied worldwide.

- **Tiered reporting.** Basic reporting for models trained with more than 10^{25} operations. Full transparency for more than 10^{26} operations. The Pillar III authority can adjust both thresholds as training efficiency changes.
- **Deployment triggers.** Developers with more than \$100 million (or equivalent) in annual AI-related revenue, or models made available to more than one million users in any one country, or more than five million users worldwide (proposed).
- **Exemptions.** Open-weight models below the compute threshold, academic research, and fine-tunes that stay below a defined additional-compute ceiling.
- **What reports cover.** Training methods, known limitations, and refusal behaviour on chemical, biological, radiological and nuclear uplift, cyber-offence, child sexual abuse material, self-harm and known jailbreaks. Discrimination testing under each country's own anti-discrimination law.
- **One common reporting format,** published by the Pillar III authority, stating what developers must test, how they document it, and what they must disclose.
- **Certification.** The Pillar III authority certifies systems at the full tier before release and can order a halt. From 2029 (proposed), certification is a condition of admission to the Protected Internet.
- **Speech.** The framework does not regulate model speech or viewpoint, and does not require any particular answer to contested political, social or moral questions. It requires only that developers disclose what safety testing they performed and what the results showed [S52].

Milestones (proposed). 2027: common reporting format. 2028: first reports. 2029: certification required for admission to the Protected Internet.

What it does not do. It does not bind developers who refuse to report; that is why Pillars I to III exist. It may miss capabilities that models hide under test [S5].

Protect the Person

Principle. No manipulative design, private by default, and no AI reasoning over a person's identifiable history, wherever that person lives.

Mechanism. Carried from v0.10 Pillars 2 and 3 [S52], applied worldwide.

- **Engagement disclosure.** Providers disclose when a system is designed to maximise engagement.
- **Wellbeing features**, such as usage notices and session limits.
- **A ban on manipulative interfaces** that trick or push people into choices they would not otherwise make and that may harm them.
- **A heightened duty toward children** and other vulnerable users.
- **Verifiable anonymisation** of retained interaction data, so that individual inputs and conversation histories cannot be traced by any system, model or human operator.
- **No reasoning over identifiable history.** Models are prevented by design from accessing, retrieving or reasoning over identifiable user history beyond the active session.
- **No profiling without a court.** No platform, partner, public authority or third party may use AI interaction data to identify, profile, target or act against a person without a court order under that country's due-process standard.
- **Link to Pillar I.** Identifiable user data is never carried on the protected tier.

What it does not do. It does not regulate content or viewpoint, and it does not reach systems outside participating jurisdictions.

Work and Livelihood

Principle. When AI displaces work at scale, the costs and gains are shared.

Mechanism. A model standard that countries can adopt, carried from v0.10 Pillar 4 [S52].

- **Coverage.** Employers with more than 1,000 employees in a country, making reductions above the greater of 10% of the workforce or 500 workers in a rolling 12 months.
- **Certification of economic necessity,** using audited criteria: declining revenue over the last four quarters, negative or materially worsening operating margin, debt-covenant pressure, or comparable filed financial indicators.
- **Presumption.** Reductions made while profits are healthy and rising are presumed to be opportunistic. The employer can rebut the presumption.
- **Transition support.** Incentives for employers that retrain their workforce to work alongside AI rather than replace it, and income support updated for longer periods between jobs, with benefits that follow workers from one employer to the next.
- **Link to Pillar II.** Manual continuity needs trained, paid people. Readiness funding includes jobs for the people who keep essential services running by hand, and for teaching those skills again.

Why. When the largest employers cut their workforces at the same time while profits are healthy, workers lose income faster than new jobs replace it, and demand falls. Workers are also the customers who sustain the economy. AI makes this more urgent, but the principle applies whatever the cause [S52].

What it does not do. It does not stop automation. It governs how large employers use it while they are profitable.

The Kill Switch readiness programme

All dates are proposed. No government has adopted them.

Phase 0, 2027. Charter and map

- A Global Readiness Coordination Centre, on the model of the International Y2K Cooperation Center, which linked coordinators from "over 170 countries" [S17].
- Each participating country names a readiness coordinator. The US government audit office credited Y2K success there in part to coordination by the national government, with periodic reporting and oversight, and issued guides on business continuity and contingency planning [S16]. The programme uses the same tools: named leads, periodic reporting, and published guides.
- Each country publishes a dependency map: what stops, and who is at risk, if shared networks and computing stop, including what each Tier 0 function depends on.
- A first global cost estimate, published and open to challenge.

Phase 1, 2027 to 2031. Protect the vulnerable first

- Voluntary, privacy-protected registries of people on home life support, with backup power, safe-use guidance for generators, and human check-in plans.
- Paper downtime procedures in every hospital, building on existing guidance [S24].
- Manual operation of water systems, with trained staff [S25].
- Black-start testing on every grid at least every three years, on the model already required in North America [S23].
- Household readiness for 72 hours [S28], moving toward one week [S26].
- Offline payment for essentials [S27].
- Funded help for lower-income countries. The International Y2K Cooperation Center reports "more than \$100 million" in direct Y2K assistance and "\$169 million" in World Bank Y2K loans [S17].

Phase 2, 2029 to 2035. World Offline Day, Rungs 1 to 3 (Annex B).

Phase 3, 2031 to 2036. Build the staged switch. Every stage reversible, and rehearsed before it is certified.

- (a) Separate the Protected Internet from the open internet at defined points.
- (b) Halt certified frontier compute through renewable on-chip operating licences, as proposed by CNAS, which estimates "as little as 18 months of involved technical effort (and up to 4 years)" to harden the hardware [S47]. CNAS proposes a licence key held by the manufacturer or a regulator. This framework requires that key to sit under Pillar III threshold custody, a multi-party arrangement CNAS did not assess.
- (c) Regional withdrawal of routing and naming. RAND judges physical disconnection of exchange points and cables "a less viable approach" than targeting routing (BGP) and naming (DNS), while calling BGP targeting "highly technically challenging" [S13]. Egypt showed in 2011 that coordinated route withdrawal can take most of a country offline, with its five major providers shutting down "an average of about three minutes apart" [S37]. That is also the abuse Pillar III guards against.

Phase 4, 2036 to 2038. Certified readiness

- Regional isolation exercises (Rung 3) in every region.
- Readiness scores for each country: how long each essential service runs with no shared network, and then with no general-purpose computing.
- The kill switch is certified only when a quorum of regions meets the floor. Until then, only the partial measures in failure mode H are available: halting certified frontier compute and isolating the protected tier as far as it exists. Shared networks are not switched off, and Tier 0 stays on.

Phase 5, 2038 to 2047 and beyond. Shrink Tier 0

- For each Tier 0 function, a funded research and engineering track toward an alternative that runs without shared networks or computing: manual, mechanical, or on isolated local controls.
- A function leaves Tier 0 only after its alternative has passed drills and been certified by the country's safety authority.
- World Offline Week (Rung 4) in regions whose readiness is certified.
- The Tier 0 list for each country is published every year.
- The end state, that humanity could switch off networks and computing entirely and survive, has no date. We do not know if or when it can be reached. The programme reports honestly on how far each function has come.

Standing review. Every ten years, an independent review of whether the programme is still needed and whether it is working.

World Offline Day

- **When.** One fixed day each year, set by the coordination centre, starting in 2029. The date is open. A day in the first week of January would recall the Y2K rollover, but it falls in northern winter and southern summer, and planners need to assess the risks of taking part in extreme cold or heat.
- **The ladder.** Each rung is attempted only after the one below has worked.
 - **Rung 1. Public offline day.** Towns, schools, councils, shops and offices volunteer to run the day on paper, cash and radio. Essential services run normally.
 - **Rung 2. Sector drills.** Hospitals, water utilities, grid operators and payment networks switch a defined part of their operations to rehearsed manual mode for a set window.
 - **Rung 3. Regional isolation exercise.** Announced in advance, after years at Rungs 1 and 2: a region separates its protected tier from the open internet, runs on it, then reconnects.
 - **Rung 4. World Offline Week** (proposed: not before 2040). In regions whose readiness is certified, several days in a row with shared networks off.
- **Tier 0 at every rung.** Tier 0 functions (Pillar II) are never switched off in any drill, for as long as they are on the list.
- **Safety rules.** Rungs 2 to 4 need the approval of each participant's own safety authority, and published stop rules that end the drill at once if safety is at risk.
- **Key ceremonies.** The Pillar III key holders rehearse meeting and acting without shared networks on every World Offline Day.
- **Reporting.** Each country publishes what worked, what failed, and what it will fix.
- **Precedents and their limits.**
 - The Great ShakeOut is a voluntary annual earthquake drill; in Washington State alone, "over one million people" were expected to take part in 2025 [S31].
 - Earth Hour says it has supporters in "over 190 countries and territories", but it switches off lights only [S32].
 - NATO's Locked Shields exercise brought together "4,000 participants from 41 nations" in 2026; it is technical, not public [S30].
 - As far as our research found, no country has rehearsed a whole society going offline.

Limits and failure modes

Each layer can fail. This table lists how, and what is left when it does.

	Failure	Effect	Mitigation	What remains
A	Switching off harms people who depend on powered systems	Deaths in hospitals and homes [S20]	Tier 0 kept on, staged switch-off, manual continuity first, voluntary registries of people on home life support, backup power with safe-use guidance	Risk reduced, not removed
B	The kill switch is captured or abused	A new tool of repression [S33, S35]	Threshold keys, public ceremonies, narrow trigger, expiry, no national kill switch	Governance can still be corrupted over time
C	The threat does not need shared networks	An actor with its own power, hardware and network keeps running	Pillar IV acts on development and compute directly	The kill switch cannot stop this case
D	The protected tier is breached	Bad actors inside the trusted tier	Verification, controlled gateways, human-controlled transfers	Isolated networks have been crossed by USB drive [S39, S40]
E	The drills cause harm	Accidents on World Offline Day	Graduated ladder, Tier 0 never switched off, regulator approval and stop rules	Accidents remain possible, especially in sector drills. Each drill publishes what went wrong.
F	The internet fragments	The network splits along political lines [S36]	Open internet untouched, protected tier limited to named life-critical functions, many operators	The fragmentation objection is still debated
G	Countries do not join, or funding falls short	The kill switch lacks the assent of "multiple international actors" that it may need [S13]	Funded help for lower-income countries, as in Y2K [S17]	Participation and funding are not guaranteed
H	The kill switch is needed before it is ready	An emergency before certification in 2036 to 2038	Partial measures only: halting certified frontier compute, isolating the protected tier as far as it exists	For most of the first decade, the last layer is incomplete
I	Tier 0 cannot be shrunk	Some life-critical functions find no workable alternative to networks or computing	Research and funding per function, tested in drills, list published yearly	The kill switch may only ever cover networks, not all computing

The kill switch is the last layer, not the plan. RAND notes that a rogue AI "may find alternative pathways to communicate" [S13].

What we could not determine

- Whether any country has rehearsed a society-wide switch-off. One example would change this framework's claim of no precedent.
- Whether a protected tier can separate within minutes at national scale, and whether Tier 0 functions can be separated from public telecoms at all in many countries.
- Whether comparable readiness practice exists in Africa, Asia and Latin America. It was not researched for this edition.
- What the programme would cost. Our estimate of hundreds of billions of dollars is not a costing.

National implementation note

Each country will adopt this framework through its own institutions. v0.10 set out one worked example, for the United States [S52]. It is kept in the archived v0.10 and summarised here as one national route, not as a model others must follow:

- A national AI safety commission of seven members with staggered fourteen-year terms, removable only for cause, with no more than four members from any one party, a chair elected by its members, funding from fees on regulated frontier developers, and an inspector general reporting to the legislature.
- Enforcement against hostile development through existing export controls on advanced chips and model weights, financial sanctions, denial of access to cloud compute, and coordinated multilateral export-control regimes.
- National law references for Pillars IV and V, such as consumer protection against unfair and deceptive practices, state privacy statutes, and constitutional search-and-seizure standards for access to interaction data.

The legal doctrine behind that example is set out in v0.10 and has not been re-checked for this edition. Other national notes are invited. We will publish them as they are written, alongside this one.

Sources

Labels: Verified means observed directly and confirmed by a second, independent method. Inferred means reasoned from one opened source. Numbers are kept from the reviewed brief; unused numbers are omitted. The list also covers sources cited only on the home page.

#	Source	Label
S1	UN Security Council, meetings coverage SC/16462, 23 September 2026	Verified
S2	Security Council Report, What's In Blue, September 2026	Inferred
S3	UN Global Dialogue on AI Governance, FAQ	Inferred
S4	UN Independent International Scientific Panel on AI, FAQ	Inferred
S5	International AI Safety Report 2026	Inferred
S6	MLex, international network of AI safety institutes renamed	Inferred
S7	Global Call for AI Red Lines	Inferred
S8	Gibson Dunn, EU AI Act Omnibus agreement	Inferred (provisional agreement)
S10	Frontier AI Safety Commitments, AI Seoul Summit 2024	Verified
S11	Palisade Research, shutdown resistance, arXiv 2509.14260	Inferred
S12	The Off-Switch Game, arXiv 1611.08219	Inferred
S13	RAND, Vermeer, PEA4361-1, November 2025	Inferred
S14	US Senate Special Committee on the Year 2000 Technology Problem, final report	Verified (\$100 billion); Inferred (global range)
S15	Government Executive, November 1999	Verified
S16	US GAO, AIMD-00-290	Inferred
S17	International Y2K Cooperation Center, final report	Verified (170 countries); Inferred (aid figures)
S18	Wikipedia, Year 2000 problem	Inferred
S19	ENTSO-E expert panel final report, 28 April 2025 blackout	Verified (date, scope)
S20	Wikipedia, 2025 Iberian Peninsula blackout	Inferred (timeline; deaths as reported)
S21	US-Canada Power System Outage Task Force, 2003 blackout final report	Inferred
S22	Microsoft, helping our customers through the CrowdStrike outage	Verified
S23	NERC EOP-005-3	Inferred
S24	HHS ASPR TRACIE, hospital downtime preparedness checklist	Inferred
S25	US EPA, water sector cybersecurity incident action checklist	Inferred
S26	MSB (Sweden), Om krisen eller kriget kommer	Verified
S27	Sveriges Riksbank, offline card payments	Inferred (goal and date only)

#	Source	Label
S28	72tuntia.fi (Finland)	Verified
S30	NATO CCDCOE, Locked Shields	Inferred
S31	US National Weather Service Seattle, 2025 Great ShakeOut	Inferred
S32	Earth Hour, about us	Inferred
S33	Access Now, #KeepItOn	Inferred
S35	Human Rights Watch, Russia's 2019 law	Verified
S36	Internet Society, internet fragmentation	Inferred
S37	Scientific American, Egypt 2011	Verified
S38	Washington Technology, GovNet, December 2001	Inferred
S39	Computerworld, agent.btz	Verified
S40	ETH Zurich Center for Security Studies, Stuxnet	Verified
S41	NERC CIP-005-7	Inferred
S42	NIST SP 800-207, Zero Trust Architecture	Inferred
S43	ICANN, the key to the internet and key ceremonies	Inferred
S44	US DoD, Nuclear Matters Handbook, chapter 8	Inferred
S45	Steven Bellovin, Permissive Action Links	Verified
S46	Adi Shamir, How to Share a Secret, 1979	Inferred
S47	CNAS, Secure, Governable Chips	Inferred
S48	Center for Data Innovation, against kill switches for AI	Inferred
S49	Nvidia, No Backdoors. No Kill Switches. No Spyware.	Verified
S51	BleepingComputer, Norsk Hydro	Verified
S52	AI Polis, A Framework for Federal AI Policy, v0.10, 8 August 2026 (archived)	Verified (text dump)
S53	Wikipedia, 2024 CrowdStrike-related IT outages	Inferred

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