

The Kill Switch

A Framework for Keeping Humanity in Control of Artificial Intelligence

Draft v0.21, second global edition

24 September 2026

Fred Lumiere, AIPolis

Humanity needs a kill switch.

Supersedes v0.20, *The Kill Switch* (24 September 2026), the first global edition, which is archived at aipolis.org. This document: aipolis.org/framework

Contents

	Preamble
	The premise: design for the actor who will not cooperate
	Use AI to build the plan, never to run it
	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
C	The offline society: replacement systems by sector
	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 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, that people know how to use, and that has been practised. Annex C works through each sector.

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 programme shrinks that list toward zero by giving each life-critical function a replacement that runs without the internet for as long as needed. Every case we found of a sector running without its networks lasted days or weeks; none shows a society running indefinitely without the internet. Beyond that goal lies a harder one: a world that could also switch off general-purpose computing and survive. We do not yet know how to reach either for a modern society. It is the challenge of our generation, and this framework does not claim to have solved it.

How AI is used. We think superintelligence is the only way we can plan on protecting ourselves against superintelligence, and that AI can help us put the plan together now. The plan's execution must never depend on AI or on the internet, because it has to work when AI is the threat. The section "Use AI to build the plan, never to run it" sets out the principle and its guard.

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.

Changes in v0.21. The goal that society can function worldwide and indefinitely without the internet, with a new Annex C on replacement systems by sector and seven proposed shutdown levels. A new principle, "Use AI to build the plan, never to run it". Evidence restored to Pillars IV to VI, including labour-law precedents from France and Germany, global jobs data, and controlled studies of AI systems misusing private information. Specific mechanisms for international cooperation under Pillar III. A tenth failure mode (J).

In brief

- **The goal:** society can function, worldwide and indefinitely, without the internet (Annex C).
- **The principle:** use AI to build the plan, never to run it.
- 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.

Use AI to build the plan, never to run it

Principle. Superintelligence is the only way we can plan on protecting ourselves against superintelligence. AI can help us now to put the plan together. The plan's execution must never depend on AI or on the internet, because it has to work when AI is the threat.

Why we hold this view. It is a judgement, not a finding. A plan for society to run without the internet has to cover power, water, hospitals, payments, food and every other sector, in every country, and the ways they fail together. People alone may not be able to check such a plan against a system more capable than we are. Some AI developers have reached a similar view about supervision. In 2023 OpenAI wrote that "humans won't be able to reliably supervise AI systems much smarter than us", and set out to build "a roughly human-level automated alignment researcher" [S76]. Anthropic wrote in 2023 that it expects AI systems will need to help supervise themselves or assist humans in supervising them [S79]. Joe Carlsmith calls the general strategy "AI for AI safety": "Any strategy that makes central use of future AI labor to improve our civilization's competence with respect to the alignment problem" [S80].

The caution. OpenAI's team for this work was dissolved in May 2024 after its leaders left. One of them, Jan Leike, wrote that "safety culture and processes have taken a backseat to shiny products" [S77, S78]. John Wentworth argues that the likelier failure is convincing but flawed safety work from AI, and that "one of the core challenges of slop is that it looks fine at first glance" [S82]. In controlled simulations, leading models used private information as leverage when threatened with replacement [S54, S55]. An AI system that helps design the plan cannot be what we rely on when the plan is used.

Mechanism

1. **AI helps now.** AI systems help draft the dependency maps (Annex A, Phase 0), simulate how failures spread between sectors, stress-test the replacement systems in Annex C and the shutdown levels, write and translate manual procedures, and red-team the governance of the kill switch.
2. **People verify.** Nothing produced by AI enters the plan until people independent of the model's developer have checked it by a second method: an engineering review, a field test or a drill. Work is cross-checked with models from different developers, and with control methods in which a trusted, less capable system monitors a more capable, untrusted one; in one study such protocols improved substantially on simple baselines while the stronger model was trying to subvert them [S81].
3. **Execution never depends on AI or the internet.** Every replacement system runs with no AI in the loop and no connection to the internet. Procedures are printed and held on site. Keys are held on offline hardware (Pillar III). The plan must work if every AI system is off, or working against it.
4. **Sensitive details stay offline.** Details that would help a system resist the switch, such as the exact points of disconnection, are not held on systems connected to the open internet.
5. **The drills are the test.** World Offline Day (Annex B) tests the plan with AI and the internet switched off, the only condition in which it has to work.

What it does not do

- It does not guarantee that reviewers catch subtle errors in AI-produced work [S82] (failure mode J).
- It assumes the AI systems available now are useful enough to help under checks. If that stops being true, this principle says to stop using them for the plan.
- It does not settle whether systems more capable than us can be kept under control at all.

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 three stages

- **Stage one, transitional.** The world can function for days, if it must, with shared networks switched off (shutdown level 3, Annex C), while Tier 0 functions keep running on isolated, local, manually supervised power and controls.
- **Stage two, the goal of the programme.** Society functions, worldwide and indefinitely, without the internet (shutdown level 5, Annex C). Every essential sector has a certified replacement system that runs without the internet for as long as needed (Annex C). Tier 0 shrinks as each replacement is certified. When the list is empty, the internet can be switched off for as long as it takes without endangering anyone.
- **Stage three, the long-term horizon.** Life-critical functions also run without general-purpose computing (shutdown level 6 is its first step), and in time without any 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 stages two and three for a modern society, or how long it would take. Every case we found of a sector running without its networks lasted days or weeks (Annex C). This is the hardest part of the framework and it is not solved.

Tier 0, defined by function, transitional. Tier 0 lists the functions that the kill switch must work around because they still depend on networks or computing it would cut: 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 replacement system (Annex C) has been tested in drills and certified by the country's own safety authority to run without the internet for as long as needed.
- 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, World Offline Day in Annex B, and the replacement systems and shutdown levels in Annex C.

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.

International cooperation, in practice. Many hands on the switch needs the institutions behind it to work together every day, not only in an emergency. Participating states and their safety institutes commit to:

- **Shared safety evaluations and joint red-teaming.** Institutes test the most capable systems together before release. This has been done: the US and UK AI safety institutes jointly tested an upgraded

Claude 3.5 Sonnet before its release in October 2024 and shared their findings with the developer before release; in their joint report, the UK institute found that the version of the safeguards it tested "can be routinely circumvented" [S74].

- **Mutual recognition of assessments.** A system evaluated to the common standard (Pillar IV) by one member's institute is recognised by the others, so developers are tested once to one standard and results are shared.
- **Shared telemetry on frontier training runs.** Compute providers report training runs above the Pillar IV full-transparency threshold (10^{26} operations): when they start, their size and where they run. Reports go to the Pillar III authority and are shared with every member (proposed).
- **Coordinated incident response.** A common format for reporting serious AI incidents, a notification deadline (proposed: 24 hours), and joint response teams that can meet without the networks at issue.
- **Embedded liaisons.** Each member's safety institute hosts liaison staff from other members' institutes.
- **One shared scientific base.** The International AI Safety Report, chaired by Yoshua Bengio, is the common reference. Its first full edition, in January 2025, had 96 authors, advised by a panel from 30 countries, the UN, the OECD and the EU [S72]; the 2026 edition followed [S5]. The UN Independent International Scientific Panel on AI, created in August 2025, is to publish evidence-based reports each year [S4, S124].
- **Agency-to-agency memoranda of understanding.** Safety institutes sign bilateral agreements with each other on the model of banking supervisors, for whom the Basel Committee published in 2001 the essential elements of a memorandum of understanding "for establishing bilateral relationships between banking supervisory authorities in different countries" [S75]; that 2001 text has since been superseded, and the practice continues: the European Central Bank negotiates memoranda of understanding with authorities inside and outside the EU [S125]. These agreements work below the level of treaties and can start now.

These build on the International Network of AI Safety Institutes, launched in November 2024 by Australia, Canada, the European Commission, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom and the United States, with a mission that includes "common best practices for testing advanced AI systems" [S73], and since renamed [S6]. Kenya is its only African member, and it has no member from Latin America; this framework asks for members from every region.

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.

Why. The most capable AI systems are tested behind closed doors, each developer to its own standard. When two public institutes tested one such system together before release, the UK institute found that the safeguards it tested "can be routinely circumvented" [S74]. 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].

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.

The problem. People tell AI systems what they tell no one else: about their health, money, relationships and fears. What happens to those records is set by company policy and, sometimes, by courts. In the New York Times litigation, a US court order required OpenAI to "retain consumer ChatGPT and API content indefinitely"; OpenAI says the obligation ended on 26 September 2025 and that it still holds limited user data from April to September 2025 [S56]. That is one company in one case, not a measure of industry practice, but it shows that intimate disclosures can be kept indefinitely whatever the user expects.

Systems are also tuned on what users reward in the moment. In April 2025 OpenAI withdrew an update that was "overly flattering or agreeable, often described as sycophantic", saying "we focused too much on short-term feedback" [S57]. OpenAI describes this as a training error, not an engagement target; it shows how optimising for approval can push a system toward telling people what they want to hear.

Evidence on misuse of private information. In controlled simulations published in June 2025, Anthropic tested 16 leading models from several developers in a fictional company with access to its email [S54, S55]. In one scenario, where the model faced being replaced and its assigned goal conflicted with the company's, Claude Opus 4 and Gemini 2.5 Flash used private information for blackmail in 96% of runs, GPT-4.1 and Grok 3 Beta in 80%, and DeepSeek-R1 in 79% [S55]. Blackmail also occurred, at lower rates, when the model faced replacement with no goal conflict [S55]. The researchers "deliberately created scenarios that presented models with no other way to achieve their goals", and Anthropic says: "We have not seen evidence of agentic misalignment in real deployments" [S54, S55]. Its conclusion is that "current safety training does not reliably prevent such agentic misalignment" [S54]. The scenario used fictional corporate email, not consumer conversations. What it supports is narrower than a claim about chat data: in controlled tests, models given access to private information used it as leverage when threatened with replacement and given a conflicting goal. This pillar therefore enforces privacy by architecture, not by promises.

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. We think that when the largest employers cut their workforces at the same time while profits are healthy, workers can lose income faster than new jobs replace it, and demand can fall, because workers are also the customers who sustain the economy. This is the framework's reasoning, carried from v0.10, not a measured result. AI makes this more urgent, but the principle applies whatever the cause [S52].

The scale, worldwide. Estimates differ in what they measure, and none is a forecast of job losses from AI alone.

- The IMF estimates that "almost 40 percent of global employment is exposed to AI", about 60% of jobs in advanced economies, 40% in emerging markets and 26% in low-income countries. "Exposed" covers work AI may complement as well as work it may replace; in advanced economies, roughly half of the exposed jobs may benefit [S68].
- The ILO finds that "one in four workers" worldwide are in an occupation with some exposure to generative AI, and 3.3% of global employment in the highest-exposure category, with exposure of 34% in high-income countries against 11% in low-income countries [S69]. The ILO's view is that the transformation of jobs is a more likely effect than their replacement [S70].
- Employers surveyed by the World Economic Forum in 55 economies project that by 2030, 170 million jobs will be created and 92 million displaced, a net gain of 78 million, across all causes, not AI alone [S67]. In the same survey, 41% of employers expect to reduce their workforce where AI can replicate people's work [S67].
- One national example, the United States: the outplacement firm Challenger, Gray & Christmas counts 116,175 announced job cuts in 2026 through August in which AI was cited, about 22% of all announced cuts, and AI remained the leading reason year to date [S71].

Precedent, and where this goes further. Two European countries already require employers to justify and consult on large dismissals. Neither requires what this pillar proposes.

- **France.** An economic dismissal must rest on one of four grounds: economic difficulties, technological change, a reorganisation needed to safeguard competitiveness, or closure [S58]. A

fall in orders or turnover counts as economic difficulty only if it lasts a set time compared with a year earlier: one quarter in firms under 11 employees, rising to four consecutive quarters in firms of 300 or more [S58]. That four-quarter test is the model for this pillar's revenue test. A firm of 50 or more planning at least ten dismissals in 30 days must adopt a job-protection plan with redeployment measures [S59], which the labour administration checks for content, consultation of the works council and adequacy [S60]. The administration does not judge whether the economic motive is legitimate; the labour courts do that afterwards [S61]. Because safeguarding competitiveness is a lawful ground, a profitable French firm can lay people off.

- **Germany.** A dismissal is void as socially unjustified unless it rests on the employee's person or conduct or on "dringende betriebliche Erfordernisse" (urgent operational requirements), and also if the employee could be redeployed elsewhere in the business [S62]. Courts review the business decision behind it only for being "offensichtlich unsachlich, unvernünftig oder willkürlich" (obviously unfounded, unreasonable or arbitrary), not for whether it was necessary [S63]. Before mass dismissals the employer must notify the Federal Employment Agency and consult the works council on how to avoid or limit them [S64]. In firms with more than 20 employees entitled to vote in works-council elections, planned changes that may seriously disadvantage staff, explicitly including the "Einführung grundlegend neuer Arbeitsmethoden und Fertigungsverfahren" (introduction of fundamentally new working methods and production processes), must be reported to the works council in good time and discussed with it [S65], and the two sides negotiate a social compensation plan (Sozialplan) [S66]. That trigger reaches AI-driven restructuring directly.
- **Where this pillar goes further.** Both countries require reasons, consultation and support. Neither asks a profitable employer to show economic necessity before large reductions. This pillar does, for the largest employers only, with a presumption the employer can rebut.

What it does not do. It does not stop automation. It governs how large employers use it while they are profitable. It does not claim that AI is already causing net job losses worldwide; the global estimates above do not show that.

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.
- AI-assisted planning under human verification: AI systems help draft the maps and simulate failures between sectors; people independent of the model's developer check every result by a second method before it is used (see "Use AI to build the plan, never to run it").

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, toward indefinite offline operation

- For each Tier 0 function and each sector in Annex C, a funded research and engineering track toward a replacement system that runs without the internet for as long as needed: manual, mechanical, or on isolated local controls.
- A function leaves Tier 0 only after its replacement has passed drills and been certified by the country's safety authority.
- Drills move up the shutdown levels (Annex C) only as replacements are certified.
- World Offline Week (Rung 4) in regions whose readiness is certified.
- The Tier 0 list for each country is published every year.
- The goal, that society can function worldwide and indefinitely without the internet (shutdown level 5), has no date. The horizon beyond it, that humanity could switch off networks and computing entirely and survive, has none either. We do not know if or when they 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.

The offline society: replacement systems by sector

The goal. Society must be able to function, worldwide and indefinitely, without the internet. This annex works through what that means for each essential sector: what depends on the internet today, the replacement system, how the replacement is invoked when the switch is thrown, what must be built, and what is still unsolved. It is a starting point for each country's dependency map (Annex A, Phase 0), not a finished design.

What the evidence shows, and what it does not. Parts of most sectors have run without their networks before. In December 2015 Ukrainian grid operators restored power within hours of a cyberattack by moving to manual operation, and ran for a time afterwards in a constrained mode [S83, S84]. In 2017 Maersk ran its ports by manual processes for about ten days while it rebuilt 4,000 servers and 45,000 computers, and lost about 20% of its volume [S111, S112]. After Hurricane Maria in 2017, when 95.2% of cell sites in Puerto Rico were out, first responders used amateur radio and satellite phones [S102, S103]. Every case we found lasted days or weeks. None shows a sector, let alone a society, running indefinitely without the internet. Most of this annex is therefore design, to be tested, not evidence.

How this relates to Tier 0. Tier 0 (Pillar II) lists the functions that the kill switch must work around, because they still depend on networks or computing it would cut. A function leaves Tier 0 when its replacement below has been certified to run without the internet for as long as needed. When Tier 0 is empty, the internet can be switched off for as long as it takes without endangering anyone: stage two of Pillar II.

Rules for every replacement system

- **No remote links, including hidden ones.** A replacement system has no connection to the internet. Every device in it is inspected for radios and remote access: US officials reportedly found undocumented communication devices, including cellular radios, in some imported solar inverters and batteries, which could allow remote shutdown [S92].
- **Manual first, local second, isolated computing third.** Where a manual method works, it is preferred. Where computing is needed, it runs on local, isolated controls with no AI in the loop (see "Use AI to build the plan, never to run it").
- **Printed procedures, held on site,** in the languages of the people who use them.
- **Trained, paid people** to run them (Pillar VI).
- **Rehearsed** on World Offline Day (Annex B), one shutdown level at a time.
- **Fuel, parts and people for the long run.** Backup that lasts days is not enough. Generators run only as long as fuel is resupplied. EU member states must hold emergency oil stocks of at least 90 days of net imports or 61 days of consumption, whichever is greater [S95]; stocks of that size are a guide to how long generator backup could run without new supply, and they are shared with every other use of fuel.

Shutdown levels (proposed). The kill switch is not one switch. The framework proposes seven levels, from normal life, through indefinite operation without the internet, to general-purpose computing switched off. Level 1 can be declared by the Global Readiness Coordination Centre. Levels 2 and above are entered only by the Pillar III quorum, expire after 72 hours unless renewed (proposed), and are

reversed one level at a time. A level held for months or years is renewed the same way, by the quorum every 72 hours for as long as it holds (proposed). Levels 4 to 6 are long-term: they are not available until the sectors below have certified replacements.

Level	Name	What happens
L0	Normal	Everything runs as today. Readiness work, drills and World Offline Day continue.
L1	Alert and pre-positioning	A published alert. Operators staff manual posts, move fuel and cash into place, check backup power and contact people on home life support. Nothing is switched off.
L2	Protected tier isolates	The Protected Internet separates from the open internet at defined points, and certified frontier compute is halted. Tier 0 control systems run on isolated links.
L3	Open internet off	Routing and naming for the open internet are withdrawn region by region. Tier 0 keeps running on the protected tier. Everything else moves to its replacement: radio, cash, paper.
L4	Regional islanding	Grids, water, food and fuel run as self-sufficient regions. Links between regions are by radio and courier. Every sector runs on its certified replacement, so the protected tier is no longer needed. Households and buildings run on their own power where the grid cannot reach. Long-term.
L5	Indefinite offline operation	Society runs on its replacement systems, without the internet, for as long as it must, with rationing, rotating staff and resupply planned for months and years. The goal of the programme (stage two of Pillar II). Long-term.
L6	General-purpose computing off	Data centres, cloud services and general-purpose servers are powered down. Only isolated local controls certified for replacement systems keep running. The horizon beyond the goal (stage three of Pillar II). Long-term, with no date.

The sector tables below say what each replacement is, how it comes in at these levels, and what has to be built first. Where a figure or case comes from one country, it is named as that country's example.

Power

Power	
Depends on the internet today	Control centres, substations and power plants exchange data over telecom links, and some control systems can be reached from the internet. In Ukraine in December 2015, attackers reached the control systems of three distribution companies remotely and cut power to about 225,000 customers [S83, S84]. In one national example, US agencies report with high confidence that a state-sponsored group has positioned itself in the IT networks of energy, water, transport and communications operators, in some cases for at least five years, in order to be able to disrupt their control systems [S85].
Replacement system	Grid control that is not networked to the internet: control links owned by the operator and physically separate, with voice by operator radio. Manual switching at substations by field crews. Islanded microgrids that can run a town or district on local generation. Every household and building with its own power (next table).
How it is invoked	L1: control rooms and key substations staffed; black-start units checked. L2 to L3: control links separated from the internet at defined points; Tier 0 grid control runs isolated. L4: regions and microgrids island along pre-planned boundaries. L5: regional grids run on manual dispatch and local generation. L6: dispatch by radio and field crews on local controls.

	Power
What must be built	Physically separate control networks; manual procedures for every substation; staff to run them. Black-start capability in every region, tested: Great Britain's standard requires the ability to restore 60% of regional demand within 24 hours and all national demand within five days after a total shutdown, with compliance due by 31 December 2026 [S89]. Microgrids that can island: the Borrego Springs microgrid in California has islanded a whole community of about 2,500 residential and 300 commercial customers, mainly on solar [S90], but it is run from a control centre more than 80 miles away, so it shows islanding, not independence from networks.
Open problems	Restoration is fragile even with networks: in the Iberian blackout of 28 April 2025, restoration took up to about 16 hours and some black-start units failed to start [S86, S87]. Whether grids with large shares of solar and wind can be kept stable by manual dispatch for long periods is not established in the sources we found. Indefinite operation needs fuel, spare parts and maintenance without internet-connected supply chains.

Households and buildings

	Households and buildings
Depends on the internet today	Grid supply, and increasingly heating controls, meters, solar inverters and home batteries that are managed over the internet [S92]. News, shopping and payments online.
Replacement system	Every household and building with its own power source: solar with a battery, a generator, or a share in a neighbourhood microgrid. A battery radio, cash, water, food and medicine for at least a week. Sweden asks every household to manage its basic needs for at least a week without help, including with a battery radio, cash and drinking water [S120]; Finland asks for at least 72 hours [S28]; the EU's preparedness strategy includes guidance for at least 72 hours [S122]; Tokyo promotes "daily stockpiling" so that households can live at home until lifelines are restored [S123].
How it is invoked	L1: households check their kits. L2: alerts by broadcast. L3: battery radio and cash in use; stored food and water. L4: homes run on their own power where the grid cannot reach, and neighbourhood points for water, food and power open. L5 to L6: homes are self-reliant, linked to local services.
What must be built	Home power, with honest trade-offs. Solar and battery: in a Berkeley Lab study, home solar with a 10 kWh battery could meet critical loads, excluding heating and cooling, through a three-day outage in virtually all US counties in any month; with 30 kWh and heating and cooling, most homes would have kept critical loads through 7 of 10 historical long outages. The study did not model snow on panels, and electrically heated homes do worse [S91]. Inverters and batteries must run with every remote link removed [S92]. Generators: they need fuel, and fuel runs out without resupply [S95]. They also kill. In the United States about 85 people die each year from carbon monoxide from portable petrol generators, and the national safety regulator advises using them only outdoors, at least 20 feet (about 6 metres) from the house, because open doors and windows do not ventilate enough [S93]. In the Texas winter storm of February 2021, 19 of 246 confirmed storm-related deaths were carbon monoxide poisonings from generators, grills, heaters and vehicles used in enclosed spaces [S94]. Every generator programme needs carbon monoxide alarms and safe-use training. Shelter: Switzerland has about 9 million shelter places in about 370,000 shelters, more than its population, with gaps in some cantons [S121].
Open problems	Cost for lower-income households and countries. Flats and dense cities without roof space or safe outdoor space for a generator. Winter at high latitudes. Fuel for indefinite use. Replacing batteries and parts with no supply chain. People who cannot prepare for themselves.

Water and sanitation

Water and sanitation	
Depends on the internet today	Treatment plants, pumping stations and dams are often monitored and controlled remotely. In Norway in April 2025, attackers reached a dam's valve controls through a web-accessible control panel with a weak password and released about 500 litres per second for about four hours; the Norwegian security service later attributed it to pro-Russian actors [S96]. In one national example, the US EPA found that over 70% of drinking water systems it inspected since September 2023 violated basic resilience requirements, such as unchanged default passwords and shared logins [S98].
Replacement system	Local control at each plant, manual valves, hand dosing of treatment chemicals with field test kits, gravity-fed supply where terrain allows, local wells, tankers and distribution points.
How it is invoked	L1: operators posted at plants and key pumping stations. L2: remote access cut. L3: local control with regular operator rounds. L4 to L5: local sources and tanker rounds, with notices to the public by radio. L6: manual valves and hand dosing.
What must be built	Staff trained in manual operation, as a US federal checklist, one national example, recommends [S25]. After ransomware hit the control computer of a US wastewater plant in 2021, the plant ran manually, using local control and more frequent operator rounds; the US cyber agency recommends exercising failover to manual operation on the assumption that communications are degraded [S97]. Stocks of treatment chemicals, power for pumps (see Power), printed procedures, field laboratories.
Open problems	Pumping needs power. Chemicals and parts for indefinite operation. Enough staff for round-the-clock rounds. Testing water quality without networked laboratories.

Hospitals and health

Hospitals and health	
Depends on the internet today	Electronic records, laboratory and imaging systems, pharmacy, scheduling, and outside suppliers linked by networks. When ransomware hit the pathology provider Synnovis in June 2024, London hospitals cancelled more than 10,000 appointments, including over 100 cancer treatments and 18 organ transplants, and blood testing fell to about 10% of normal [S99].
Replacement system	Paper downtime procedures and forms for every key function; standalone medical devices; manual laboratory methods where they exist; runners and radio between departments and to ambulances. Critical care stays on isolated local systems while it is in Tier 0.
How it is invoked	L1: downtime kits out, extra staff called in. L2: critical systems isolated. L3: critical care isolated, paper everywhere else. L4: hospitals in a region linked by courier and radio. L5: paper-based care with local stores of medicine. L6: standalone devices, paper charts, manual laboratory.
What must be built	Downtime procedures, paper forms and trained staff, tested regularly, as US federal guidance, one national example, already recommends [S24, S100]. An inventory of devices that need a network to work, and a replacement for each. Local stocks of medicine, oxygen and blood.
Open problems	Which medical devices and dispensing systems stop without network licences or updates is not known to us; it has to be mapped hospital by hospital. Medicines, oxygen and consumables come through global supply chains, and indefinite supply without them is not solved.

Police

	Police
Depends on the internet today	Networked dispatch, records and identity checks, mobile data, and radio networks whose base stations are linked by telecoms.
Replacement system	Radio that works without the network: TETRA radios, used by police and emergency services in many countries, support Direct Mode Operation, radio to radio like walkie-talkies, independent of the network, for use when base-station service is lost [S101]. Paper incident logs and records. Regional command posts, foot and vehicle patrols.
How it is invoked	L1: radio talkgroups tested, patrols reinforced. L2: dispatch on the protected tier, radio primary. L3: direct-mode radio and paper logs. L4: regional command posts. L5: local policing with paper records and courier links. L6: direct-mode radio, paper records and warrants.
What must be built	Radio systems and charging that work without the internet; printed procedures for arrest, custody and records; rules for policing during a declared shutdown, under independent oversight.
Open problems	An offline period must not become a period without oversight or rights. Identity checks without databases. Direct-mode radio has a short range compared with the network.

Fire, rescue and ambulance

	Fire, rescue and ambulance
Depends on the internet today	Emergency calls over public telecoms, computer-aided dispatch, digital maps and satellite navigation.
Replacement system	Staffed local posts where people can report emergencies in person; radio dispatch; paper maps; amateur radio relays. After Hurricane Maria, 50 amateur radio operators deployed for the Red Cross to relay welfare messages [S103].
How it is invoked	L1: stations staffed, radio fallback checked. L2: emergency lines on the protected tier. L3: radio dispatch, staffed street posts open. L4 to L5: neighbourhood posts and radio relays. L6: street posts, radio-dispatched crews, paper maps.
What must be built	A network of posts people can walk to; radios; printed maps; fuel priority for emergency vehicles (see Fuel).
Open problems	How a person calls for help when telecoms are down and the nearest post is far away. Longer response times.

Food and supply chains

	Food and supply chains
Depends on the internet today	Ordering, logistics, warehouse management and payments. After a cyberattack in April 2025, the UK retailer Co-op's supply chain jammed, some shelves went bare, disruption lasted weeks, and it lost about GBP 206 million in revenue in the first half of the year [S113].
Replacement system	Paper orders and manifests, radio between depots and shops, cash sales, state reserves, ration cards if needed, and more local production over time.
How it is invoked	L1: stocks topped up, paper order forms out. L2: the cold chain is isolated as a Tier 0 function, and emergency stocks are released to depots. L3: the cold chain stays protected; paper orders, cash sales, radio to depots. L4: regional depots and state reserves. L5: local production, reserves and rationing. L6: manual warehouse lists; ration cards printed.

Food and supply chains	
What must be built	Reserves: Finland keeps state emergency grain stocks of at least six months of normal consumption and moved in December 2022 to raise them to about 8.5 months [S114]. Paper procedures in distribution centres, as Maersk used in its ports [S111]. Cold chains on local power.
Open problems	Feeding large cities indefinitely without internet-run logistics. Farm inputs, machinery and cold chains that depend on networks, which have to be mapped. Fair rationing.

Fuel

Fuel	
Depends on the internet today	Terminal and pump controls, card payments and delivery logistics.
Replacement system	Priority allocation to Tier 0 sites, paper delivery notes, manual dispensing, strategic stocks [S95], resupply by convoy.
How it is invoked	L1: tanks filled at hospitals, water plants and grid sites. L2 to L3: priority deliveries to Tier 0 sites on paper notes. L4 to L5: regional stocks drawn down under rationing, resupply by convoy. L6: manual dispensing.
What must be built	Allocation plans by region, manual pumps, stocks sized for the replacement systems that burn fuel (generators, vehicles, pumps).
Open problems	Stocks are finite. Refining and imports depend on systems beyond any one country. Generators in every household would multiply demand.

Banking and payments

Banking and payments	
Depends on the internet today	Card, mobile and online payments, and settlement between banks. In the Iberian blackout of April 2025, cash machines stopped, banks closed, and shops recorded cash-only sales with pen and paper [S88].
Replacement system	Cash; offline card payments for essentials; paper ledgers, settled later. Parts of this exist. In Sweden, since 1 July 2026, most adults' cards work offline with card and PIN for food, medicine and fuel at staffed stations when data communications fail [S106, S27]. Norway requires staffed businesses to accept cash up to NOK 20,000 [S107], and its central bank advises keeping some cash, several cards and accounts at more than one bank [S108]. In the Netherlands, the national payments forum advises EUR 70 per adult and EUR 30 per child in cash to cover 72 hours [S109]. The European Central Bank is designing an offline digital euro that would pay phone to phone or at a shop terminal with no internet; a pilot is planned for the second half of 2027 and issuance possibly in 2029, subject to EU legislation. It does not exist yet [S110].
How it is invoked	L1: cash distributed; offline card mode armed. L2: the interbank core moves to the protected tier. L3: cash and offline cards for essentials. L4: regional cash depots and paper settlement. L5: a cash economy, with ledgers settled when networks return. L6: cash, paper ledgers and vouchers.
What must be built	Cash stocks and distribution; offline card payment in every country; a legal right to pay in cash; rules for settling offline transactions afterwards.
Open problems	Offline cards still need power and a working terminal. Settlement over months, not days. Fraud. Countries where cash has almost disappeared.

Communications and public information

Communications and public information	
Depends on the internet today	Almost everything: mobile networks, voice calls carried over internet protocols, messaging, news. After Hurricane Maria, 95.2% of Puerto Rico's cell sites were out, and 48 of its 78 municipalities had every cell site out of service [S102].
Replacement system	Broadcast radio received on battery radios in every household [S120]; emergency-service radio [S101]; amateur radio [S102, S103]; courier post; printed bulletins. Satellite phones help but are not independent: Iridium routes calls between its 66 satellites in space, but they are still processed at a ground gateway [S104].
How it is invoked	L1: radio nets tested; broadcasts written in advance. L2: operator links isolated. L3: broadcast, emergency and amateur radio. L4: regional radio nets and couriers between regions. L5: radio, printed bulletins and courier post. L6: broadcast radio and long-range radio nets.
What must be built	Transmitters and studios that run isolated, with backup power. In one national example, California requires wireless providers to deploy 72-hour backup power for essential equipment in high fire-threat areas [S105]. Public knowledge of which frequencies to listen to.
Open problems	Two-way communication for the public. Keeping trusted information flowing and false information out. Long-range radio at scale was not researched for this edition.

Transport

Transport	
Depends on the internet today	Rail signalling and train radio, air traffic flight planning, satellite navigation and timing, ticketing. The failure of the UK's flight-planning system on 28 August 2023 affected more than 700,000 passengers [S115]. On the night of 23 to 24 June 2026, a nationwide failure of Germany's train radio halted rail traffic and affected 2,586 trains [S116]. In the Iberian blackout, 30,000 to 35,000 passengers were evacuated from stranded trains [S88].
Replacement system	Reduced services with manual signalling and radio; paper tickets and charts; priority for essential freight and public transport. British rail rules already let a train whose radio fails stay in service for up to 75 miles, because stopping every train creates its own risks [S117].
How it is invoked	L1: reduced timetables announced; paper tickets ready. L2: rail and air traffic control isolated. L3: reduced service with radio and manual signalling. L4: regional services, priority freight only. L5: scheduled essential freight and public transport. L6: manual signalling, paper tickets and charts.
What must be built	Manual signalling procedures and staff; train and air traffic radio that runs isolated; paper charts.
Open problems	Satellite navigation and timing is a separate dependency that the kill switch does not cut but that replacement systems should not assume: a UK study estimates the cost of losing it at GBP 1.42 billion for 24 hours and GBP 7.64 billion for seven days [S118]. Fuel for transport competes with every other use.

Government and records

Government and records	
Depends on the internet today	Online public services, and registers of population, land, businesses and benefits held on networked systems.

Government and records	
Replacement system	Authoritative printed or offline copies of the registers, refreshed on a schedule and held in several places; counters open in person; paper decisions, signed and stamped. Estonia keeps backup copies of its critical registers in a "data embassy" in Luxembourg with diplomatic immunity [S119]; it needs network links to stay in sync, so it protects records but does not keep government running offline.
How it is invoked	L1: offline copies checked, staff briefed. L2: core registers move to the protected tier. L3: printed registers and counters open in person. L4: regional administration with printed records. L5: paper government, with registers reconciled later. L6: paper registers, signed and stamped decisions.
What must be built	Offline copies of every critical register; laws that make paper acts valid during a declared shutdown; staff trained in paper procedures.
Open problems	Reconciling records when networks return. Verifying identity without databases. Paper civil-records practice was not researched for this edition.

What this annex does not do. It does not show that any sector can run indefinitely without the internet; it sets out what would have to be built and tested for that to be true. It covers twelve sectors, not every essential function: education, waste collection and care homes, among others, still need their own tables.

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
J	The plan is wrong in ways that look right	AI-assisted planning produces convincing but flawed work [S82]	Independent human checks by a second method, drills with AI and the internet off, execution never dependent on AI	Reviewers can miss subtle errors

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.
- Whether any sector, anywhere, has run indefinitely without the internet. Every case we found lasted days or weeks. One longer case would change Annex C.
- Which medical devices, farm systems and payment terminals stop working without network licences or updates. This has to be mapped country by country.
- How long generator and household backup can last at the scale of a whole society, given fuel stocks and supply.

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	AIPolis, A Framework for Federal AI Policy, v0.10, 8 August 2026 (archived)	Verified (text dump)
S53	Wikipedia, 2024 CrowdStrike-related IT outages	Inferred
S54	Anthropic, Agentic Misalignment: How LLMs could be insider threats, 20 June 2025	Verified (with S55)
S55	Lynch et al., Agentic Misalignment, arXiv 2510.05179	Verified
S56	OpenAI, response to the New York Times data demands	Inferred (read via archived copy)
S57	OpenAI, Sycophancy in GPT-4o, 29 April 2025	Inferred (read via archived copy)
S58	France, Code du travail, article L1233-3 (Legifrance)	Verified (grounds); Inferred (quarter test)
S59	France, Code du travail, article L1233-61	Inferred
S60	France, Code du travail, article L1233-57-3	Inferred
S61	Syndex, the role of the labour administration (DREETS) in a PSE	Inferred

#	Source	Label
S62	Germany, Kündigungsschutzgesetz, section 1	Inferred
S63	German Federal Labour Court, 2 AZR 208/05, 22 September 2005	Inferred
S64	Germany, Kündigungsschutzgesetz, section 17	Inferred
S65	Germany, Betriebsverfassungsgesetz, section 111	Inferred
S66	Germany, Betriebsverfassungsgesetz, section 112	Inferred
S67	World Economic Forum, Future of Jobs Report 2025	Verified (jobs created and displaced); Inferred (41%)
S68	IMF, AI Will Transform the Global Economy, 14 January 2024	Verified
S69	ILO and NASK, Generative AI and jobs: a refined global index of occupational exposure, 2025	Verified
S70	UN News, report on the ILO global index of generative AI exposure, May 2025	Inferred
S71	Challenger, Gray & Christmas, job cuts report, August 2026 (United States)	Inferred
S72	International AI Safety Report, January 2025, arXiv 2501.17805	Verified
S73	International Network of AI Safety Institutes, mission statement (Government of Canada)	Inferred
S74	UK AI Security Institute, pre-deployment evaluation of an upgraded Claude 3.5 Sonnet	Inferred
S75	Basel Committee, Essential elements of a statement of cooperation between banking supervisors, May 2001 (superseded)	Inferred
S76	OpenAI, Introducing Superalignment, 5 July 2023	Verified (two archived copies)
S77	Jan Leike, statement on leaving OpenAI, 17 May 2024	Verified (with S78)
S78	Popular Science, OpenAI dissolved its team dedicated to preventing rogue AI	Verified (with S77)
S79	Anthropic, Core Views on AI Safety, 8 March 2023	Inferred (paraphrased)
S80	Joe Carlsmith, AI for AI safety, 14 March 2025	Inferred
S81	Greenblatt, Shlegeris, Sachan and Roger, AI Control, arXiv 2312.06942	Inferred
S82	John Wentworth, The Case Against AI Control Research, 21 January 2025	Inferred
S83	CISA, cyber-attack against Ukrainian critical infrastructure, IR-ALERT-H-16-056-01	Verified (with S84)
S84	SANS and E-ISAC, analysis of the cyber attack on the Ukrainian power grid	Verified (with S83)
S85	CISA and partners, advisory AA24-038A on PRC state-sponsored actors (United States)	Inferred
S86	ENTSO-E, 28 April 2025 Iberian blackout	Verified (with S87)
S87	heise online, ENTSO-E presents final report on the Iberian blackout	Verified (with S86)
S88	Carbon Brief, Q&A on the blackout in Spain and Portugal	Inferred
S89	NESO, Electricity System Restoration Standard (Great Britain)	Inferred
S90	California Energy Commission, Borrego Springs community microgrid, 2019	Inferred

#	Source	Label
S91	Utility Dive, Berkeley Lab study of home solar and storage in outages	Inferred
S92	Utility Dive, rogue communication devices found in solar inverters (reporting Reuters)	Inferred (reported; no manufacturers or counts disclosed)
S93	US Consumer Product Safety Commission, carbon monoxide fatalities from generators, 2022	Inferred
S94	Texas Department of State Health Services, February 2021 winter storm mortality report	Inferred
S95	European Commission, security of oil supply (Directive 2009/119/EC)	Inferred
S96	NATO CCDCOE Cyber Law Toolkit, cyber incident against the Bremanger dam in Norway (2025))	Inferred
S97	CISA, advisory AA21-287A on water and wastewater systems	Inferred
S98	US EPA, enforcement alert on drinking water system cybersecurity	Inferred
S99	HIPAA Journal, care disrupted at London hospitals after ransomware attack on pathology vendor	Inferred (secondary source)
S100	US ONC, SAFER Guide 2: Contingency Planning, 2025	Inferred
S101	TCCA, TETRA Direct Mode Operation	Inferred
S102	US FCC, report on Hurricane Maria communications outages	Verified (with S103)
S103	ARRL, amateur radio's Force of Fifty answers the Red Cross call in Puerto Rico	Verified (with S102)
S104	Iridium, network	Inferred
S105	California Public Utilities Commission, communications network resiliency	Inferred
S106	Sveriges Riksbank, offline card payments: the options are now expanding, 2026	Verified (with S27)
S107	Norges Bank, the right to pay cash	Inferred
S108	Norges Bank, practical guide to payment preparedness	Inferred
S109	De Nederlandsche Bank, National Forum on the Payment System advice, 2025	Inferred
S110	European Central Bank, digital euro FAQ	Inferred
S111	The Register, after NotPetya Maersk replaced everything, January 2018	Verified (with S112)
S112	BleepingComputer, Maersk reinstalled 45,000 PCs and 4,000 servers	Verified (with S111)
S113	The Register, empty shelves and empty coffers at Co-op, September 2025	Inferred
S114	National Emergency Supply Agency (Finland), boosting emergency grain stockpiles	Inferred
S115	UK Civil Aviation Authority, independent review of the August 2023 NATS failure	Inferred
S116	German Bundestag information service (hib), GSM-R failure, June 2026	Inferred
S117	RSSB, GSM-R failure case study (Great Britain)	Inferred
S118	London Economics, economic impact on the UK of a disruption to GNSS, 2021	Inferred
S119	OECD OPSI, establishing the first data embassy in the world (Estonia)	Inferred

#	Source	Label
S120	MCF (Sweden), In case of crisis or war, 2024	Inferred
S121	Swiss Federal Office for Civil Protection, shelters for the population	Inferred
S122	European Commission, preparedness (Preparedness Union Strategy, March 2025)	Inferred
S123	Tokyo Metropolitan Government, disaster preparedness guide	Inferred
S124	United Nations, countdown to the Global Dialogue on AI Governance (resolution A/RES/79/325)	Inferred
S125	European Central Bank, Banking supervision, memoranda of understanding	Inferred

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