CLTC
Center for Long-Term Cybersecurity
UC Berkeley

CNA
ANALYSIS & SOLUTIONS

CYBERSECURITY FUTURES 2025
INSIGHTS AND FINDINGS

FEBRUARY
2019

Project Preface

Cybersecurity Futures 2025 is a collaboration between the University of California, Berkeley Center for Long-Term Cybersecurity (CLTC) and CNA's Institute for Public Research, conducted in partnership with the World Economic Forum's Global Future Council on Cybersecurity (2016-2018) and the Forum's Centre for Cybersecurity.

This report includes a short description of the process and evolution of the project, along with summary insights from workshops conducted around the world in 2018. It also includes the four scenario narratives that were the foundation for the workshops.

The project website, cyberfutures2025.org, features a set of short videos that narrate the key elements of the four scenarios, along with an introductory video that situates these stories and explains how to use them. The site also includes a tool that invites personal interaction with the scenarios and provides a heuristic to inform strategic decision-making.

We hope that readers of this report and the accompanying multimedia content will uncover insights to help drive their organizations to be more anticipatory, more proactive, and ultimately more successful in addressing a wide range of emerging cybersecurity challenges. We welcome your feedback as you interact with these ideas.

We would like to thank the World Economic Forum and its Centre for Cybersecurity for their collaboration throughout this process. We would also like to thank the organizations that helped support this work, including HP, Inc.; Symantec; Qualcomm; and CyberCube, as well as the entities that hosted our workshops. Most importantly, we would like to thank the many colleagues who contributed ideas and critiques to the process of creating the scenarios, as well as the community of experts from industry, government, and civil society who participated on the Global Future Council on Cybersecurity, attended and contributed to our workshops, and helped derive and synthesize the insights from this process.

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Steven Weber
Professor, UC Berkeley School of Information
Director, Center for Long-Term Cybersecurity

CNA
ANALYSIS & SOLUTIONS

David Kaufman
Vice President and Director
Dawn Thomas
Associate Director
CNA Institute for Public Research

Steptoe
STEPTOE & JOHNSON LLP

Alan Cohn
Partner, Steptoe & Johnson LLP
Adjunct Professor, Georgetown University Law Center
Co-Chair, World Economic Forum Global Future Council on Cybersecurity (2016-2018)

Contents

Project Description and Summary Insights

BACKGROUND AND OVERVIEW

One observation consistently made about the digital era is that when people and technology mix, the results are surprisingly hard to anticipate. This kind of uncertainty puts cybersecurity professionals at a structural disadvantage because it favors attackers over defenders and protectors. Looking to the future, at the intersection of people and digital technology, there is a gulf between the operational security on the agenda today and the range of cybersecurity issues and challenges that will emerge in a decision-relevant future time frame.

To address this gap, we developed a set of future-looking cybersecurity scenarios that are intended to spur a much-needed discussion about the cybersecurity challenges that government, industry, and civil society will face in the future, beyond the immediate horizon.

Cybersecurity Futures 2025 rests on the foundational idea that if we can anticipate how cybersecurity challenges will evolve and understand how governments, firms, and societies in different parts of the world think about those challenges, we can better position decision-makers to reduce detrimental frictions and seize opportunities for cooperation. By tapping into a broadly felt sense that current policy and strategy frameworks in cybersecurity are inadequate and becoming more so, Cybersecurity Futures 2025 seeks to provide a roadmap for new high-level concepts and strategies that drive operational and tactical adaptation in the future.

PHASE 1: DEVELOPING THE CYBERSECURITY FUTURES 2025 SCENARIOS

In the first phase of the project, we developed a set of scenarios that portray a possibility space of "cybersecurity futures" looking forward to roughly 2025. These four scenarios were designed to stress trade-offs in goals and values that will appear in the near future. The scenarios focus on what is relevant and plausible, while also challenging existing beliefs. They were specifically designed to elicit meaningfully different points of view from different parts of the world.

The Cybersecurity Futures 2025 scenarios (like all scenarios) are not predictions. They are logical narratives that tell stories about how forces of change from a variety of sources—technology, economics, human behavior, corporate strategy, government policy, social and ethical dimensions, and more—may overlap and combine to create a set of cybersecurity problems in 2025 that are different from those encountered today. This future problem set involves a broader set of actors, has greater stakes, sits on different technological foundations, and engages core human values in a novel way. The four scenarios are attached as an appendix to this summary note.

PHASE 2: INTERNATIONAL WORKSHOPS

Between May and October 2018, we took these scenarios to seven international locations: Palo Alto, Munich, Singapore, Hong Kong, Moscow, Geneva, and Washington, DC. In each location, we organized a workshop with a mix of participants from government, business, civil society, academia, and other domains. We ran similar workshop processes in order to extract reactions and insights that would be roughly comparable. These comparisons are the most important immediate product of the workshops. Though none of the four scenarios will "come true" in 2025, it is very likely that cybersecurity in 2025 will encompass many of the issues and challenges that these scenarios portray. Anticipating reactions in different parts of the world contributes to a forward-looking research and policy agenda that should be more robust, intellectually and practically—and more broadly applicable across countries and regions.

PHASE 3: GENERATING INSIGHTS

A set of summary insights took shape from the results of the seven workshops. These insights come with obvious caveats, the most important of which is the use of aggregate geographical categories as placeholders. Ascribing the outcomes of a workshop in Munich to "Europe," for example (despite broad representation from a number of European countries, institutions, and sectors), is not the same as holding workshops across Europe, or dividing perspectives among the various countries and regions of Europe. The geographic labels are best thought of as imperfect proxies and conceptual "clouds" with fuzzy edges. Another caveat is recency bias; our workshop participants are people, and people read future scenarios in the context of what is most important and urgent in their minds at that moment. We designed our workshop process to minimize these kinds of biases, but it is impossible to fully eliminate them.

Caveats notwithstanding, we believe that the early insights we report below are at least directionally correct and, thus, deserve focused attention in strategic planning and future decision-making. We offer three overarching observations, and propose five new landscape elements that reframe the decision-making environment.

Overarching Observations

1. It is notable that the discourse about digital technology and security is now deeply "nationalized" and has become even more so in the context of our scenarios. As recently as three years ago, a "free and open internet" narrative that placed governments squarely in the background of the digital environment was still robust. That ideology, which in some respects was naive, appears to be largely gone. "Data nationalism" of some kind is now a given. The new narrative centers on technology firmly yoked to the goals of national power. While this is more historically familiar, it is also a significant discontinuity for the internet and the digital economy.
2. There is a strong sense of disillusionment with vague discussions about "cyber-norms." Workshop participants around the world were hard-pressed to attach concrete meaning to norms, or to articulate how discussions about norms would

lead—as opposed to follow—emergent behaviors.

3. Some of the most profound upside expectations about what digital technology could do to improve the human experience risk becoming buried in the emerging landscape. The first generations of digital technology came with (possibly outsized) idealism—for wealth creation, safety, efficiency, peace, happiness and more. It was inevitable that those expectations would be adjusted over time. But if the pendulum swings too fast and too far towards the pole of risk and threat—as now appears possible—societies risk losing sight of the massive good these technologies can do if properly managed and secured.

New Landscapes

1. The "golden mean" of light-touch regulation and permission-less innovation that governments and business have carved out together as a foundation for the digital economy over the past 20 years is not necessarily enduring. In our workshops, participants did not try to rescue some version of this formula—by which companies have the freedom to develop and deploy new technologies unless it is shown definitively that those technologies are dangerous—because it was not visible to them how it could become an effective route to improved digital security. The idea that this formula is broken, even as an aspiration, is a significant change in the political-economic environment, and we should expect diverging experiments in new regulatory regimes around the world. While those experiments will share a greater role for governments overall, the global landscape will become increasingly variegated.

   This provokes a simple question: Who should lead the charge to course-correct if (perhaps when) things go wrong? In Palo Alto, the answer was "It will have to be the large firms since that is where the capability lies." In Munich, it was "Europe lacks the firms, and we do not trust governments to respond, so we need a citizen social movement." In Singapore, the reaction was more muted: "It probably will not go that wrong, but if it does, the government is the fixer-of-last-resort." Those are very different trajectories that would grate against each other in important ways.

2. Digital geopolitics is no longer a layer superimposed on conventional geopolitics; digital is creating new alignments among new actors, and not only states. At present, there are many who retain the belief that "no one really goes to war over a cyber-attack and, if they do, it is not really about the cyber-attack per se." Our workshops suggest this belief will not endure. Alliances are being reshuffled: arguments about cyber-attack attribution in Europe, for example, focus as much on the US NSA as they do on groups such as APT-28. Parastatal and criminal organizations are becoming equal-status players to large firms and governments: to refer to them as "non-state actors," implying second-tier geopolitical status, is mistaken. Likewise, "large firms and governments" are now widely seen as nearly co-equal participants in the political

process; countries such as Denmark have already created a formal ambassador to the technology sector, and more will follow. The emergence of new technologies that could drastically reshuffle geopolitical power (possibly quantum computing, for example) will accelerate the reformulation of alliances relating to digital interests, and it is possible that firms will be as significant as states in the new alignments. There will also emerge new definitions of what constitutes criminal activity, and of who or what is a "criminal." As those definitions are diverging across geographies, the opportunities for digital criminals to arbitrage within the global marketplace will increase.

3. Digital-induced job displacement and inequality will become more than a stressor; these dynamics are set to bring fundamental breakdowns and failures in both labor markets and politics. Social capital and broader societal resilience will be critical assets in navigating the transition towards any new automation and robotics-enabled labor market equilibrium.

   Countries and regions are positioned very differently on this dimension; for example, Asians seem to hold a higher level of confidence that societies can endure through these changes, built on the belief that many Asian societies have proven to be resilient and cohesive in the face of comparable challenges. However, there is also a looming recognition that economic growth and development trajectories for most countries are increasingly uncertain. Populist movements in the US and Europe demonstrate in part the strains resulting from a loss of confidence that a mix of conventional markets and politics will ensure the benefits of digital technology help those seemingly being left behind. The success story of the late industrial-era developing country (low-wage manufacturing evolves towards higher value-add along with capital accumulation) is now largely obsolete and the path for late developers to succeed in a global economy dominated by data flows and machine learning has not been defined. Transnational movements—either of distressed and displaced labor, or perhaps of the (massively empowered) technology elite labor force—are nascent in some parts of the world (particularly the US and aspirational in Europe); their possible emergence would become an important new part of the security landscape.

4. The largest intermediation platform firms are now seen everywhere as a truly distinct category of player, whose relationships with governments, consumers, and societies need special assessment, attention, and, possibly, oversight. A striking observation is that while many of the platforms are global, or becoming so, conversations about their societal and economic consequences remained national or regional at best. Market power and oligopoly is now an assumption in most of the world; Europeans emphasize the negative implications most strongly. In Asia, the emphasis falls on speech, and how the act of trying to assess "truth" in platform-structured discourse affects social capital and cohesion. America struggles with the consumer-welfare focus of US competition policy; there is little visibility into (and relatively little concern about) how US-based platform firms affect societies and economies outside the US.

These contemporary observations remained largely robust in the context of the 2025 scenarios, though changes in computing architecture were seen as destabilizing. What is clear is that competition policy and cybersecurity policy are converging in many respects, and this trend brings national differences in approaches to competition policy into the security landscape as well.

5. The cybersecurity challenge of protecting networks and datasets from sovereign and criminal thieves is morphing into a challenge of protection from devious manipulation. Brute force attacks remain on the agenda, but there is a broad assumption that the sophistication of attacks is set to rise through some of these more insidious channels, such as adversarial machine learning, subtle deep fakes, or small changes in training set data that intentionally bias algorithms. This will accelerate the trend of cybersecurity becoming a much more scientifically interesting area, but it will also pile even more demand on a workforce that is already under massive stress. Broad societal resilience programs are one response that is talked about more in Asia than elsewhere; in the US, consumers and users are still seen as mostly passive, and the concept that there is an ability to educate them to be savvier consumers of information is still nascent. Turning more of the burden over to automated systems such as artificial intelligence-driven platforms may be another credible response—with substantial differences in what roles and controls should be maintained for human decision-making.

PHASE 4: WHAT'S NEXT?

As a result of these observations, we believe that senior decision-makers developing cybersecurity strategies in government and the private sector must now engage with each of the following questions, individually and collectively, on an ongoing basis over the next few years. These are obviously not operational-level questions specific to a particular industry sector or country. However, the answers to and hypotheses on these questions should inform operational plans that are more robust in a fast-evolving environment.

- Where are the new deviant digital black markets evolving? And what is being traded in those markets?
- What is the definition of a criminal? And what are the arbitrage-ready differences among those definitions?
- What new geopolitical alliances are forming and emerging? And how could we better understand the granular nuances of interest cleavage within nations and societies that influence the direction those alliances might take?
- How much digitally exacerbated and/or induced inequality can different societies absorb? And at what rate?
- Where are first-mover advantages to be found—in technologies, of course, but also in policies?

- What characteristics make a society resistant and resilient to digital manipulation? If employees, consumers and citizens need to be reoriented as less-passive players in the cybersecurity landscape of 2025, what new capabilities do they need to attain and how can they attain them?

Grappling with these questions should be a defining focus in 2019 for the C-suite, boards, and government agencies in essentially every country around the world.

Scenario Summaries

Scenario 1—Quantum Leap

The year is 2025, and the first countries to achieve practical quantum computing capabilities have spent the past several years trying to construct a non-proliferation regime that would preserve the economic, strategic and military advantages the technology has begun to generate. But other countries—and even large cities—that are behind in the race have resisted the offer to access watered-down quantum services from the few elite providers in return for restraint in development. Instead, many attempt to pursue "quantum autonomy". Technology development accelerates almost to the exclusion of ethical, economic and other sociopolitical concerns as quantum leaks into the "deviant globalization" sphere of drug cartels and other worldwide criminal networks. Ultimately, the carrots of a restrictive non-proliferation bargain aimed at governments have not been enticing enough (and the sticks not fearsome enough) to hold a regime together, and the model that more or less worked to contain the spread of nuclear weapons in a previous era fails with quantum. In 2025, the Americans and the Chinese in particular are starting to wonder if their next best move is to reverse course and speed up the dissemination of quantum computing to their respective friends and allies, while the deviant sector is racing ahead.

Scenario 2—The New Wiggle Room

This is a world in which the promise of secure digital technology, the Internet of Things (IoT) and large-scale machine learning (ML)—to transform a range of previously messy human phenomena into precise metrics and predictive algorithms—turns out to be in many respects a poisoned chalice. The fundamental reason is the loss of "wiggle room" in human and social life. In the 2020s, societies confront a problem opposite to the one with which they have grappled for centuries: now, instead of not knowing enough and struggling with imprecision about the world, we know too much, and we know it too accurately. Security has improved to the point where many important digital systems can operate with extremely high confidence, and this creates a new set of dilemmas as precision knowledge takes away the valuable lubricants that made social and economic life manageable. As the costs mount of not being able to look the other way from uncomfortable truths, or make constructively ambiguous agreements, or agree to disagree about "facts" without having to say so, people find themselves seeking a new source of wiggle room. They find it in the manipulation of identity—or multiple and fluid identities. This effort to subtly reintroduce constructive uncertainty and recreate wiggle room overlaps with the emergence of new security concerns and changing competitive dynamics among countries.

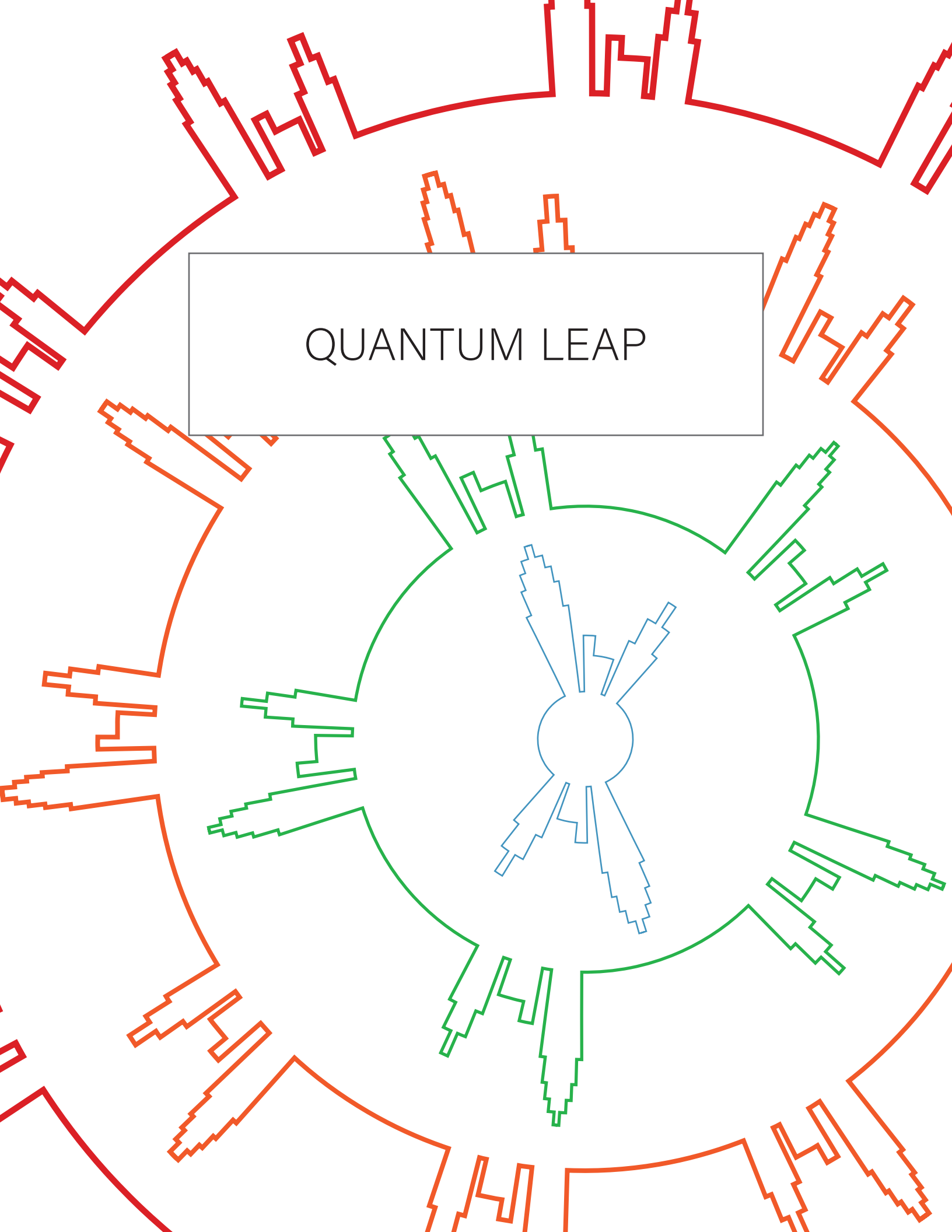
Scenario 3—Barlow's Revenge

As digital security deteriorates dramatically at the end of the 2010s, a broad coalition of firms and people around the world come to a shared recognition that the patchwork quilt of governments, firms, engineering standards bodies and others that had evolved to try to regulate digital society during the previous decade was no longer tenable. But while there was consensus that partial measures, piecemeal reforms and marginal modifications were not a viable path forward, there was also radical disagreement

on what a comprehensive reformulation should look like. Two very different pathways emerged. In some parts of the world, governments have essentially removed themselves from the game and ceded the playing field for the largest firms to manage. This felt like an ironic reprise of the 1996 ideological manifesto of John Perry Barlow, "A Declaration of the Independence in Cyberspace". In other parts of the world, governments have taken the opposite path and embraced a full-bore internet nationalism in which digital power is treated unabashedly as a source and objective of state power. In 2025, it is at the overlaps and intersections between these two self-consciously distinctive models, existing almost on different planes, that the most challenging tensions but also surprising similarities are emerging.

Scenario 4—Trust Us

This is a world in which digital insecurity in the late 2010s brings the internet economy close to the brink of collapse, and in doing so, drives companies to take the dramatic step of offloading security functions to an artificial intelligence (AI) mesh network, "SafetyNet", that is capable of detecting anomalies and intrusions, and patching systems without humans in the loop. Fears that AI would disrupt labor markets are turned on their head as the AI network actually helps the economy claw its way back from the brink, and restores a sense of stability to digital life. But a new class of vulnerabilities is introduced, and while SafetyNet is for many purposes a much less risky place, the security of the AI itself is consistently questioned. In 2025, most people experience the digital environment as a fractured space: an insecure and unreliable internet, and a highly secured but constantly surveilled SafetyNet organized and protected by algorithms. Institutions can breathe a little easier as they segregate their activities into either environment. But many individuals are wondering whether the features of reality that matter to them—the values they see as worth securing—have been trampled along the way.


QUANTUM LEAP

Quantum Leap

The year is 2025, and the first countries to achieve practical quantum computing capabilities have spent the past several years trying to construct a non-proliferation regime that would preserve the economic, strategic and military advantages the technology has begun to generate. But other countries—and even large cities—that are behind in the race have resisted the offer to access watered-down quantum services from the few elite providers in return for restraint in development. Instead, many attempt to pursue "quantum autonomy." Technology development accelerates almost to the exclusion of ethical, economic and other sociopolitical concerns as quantum leaks into the "deviant globalization" sphere of drug cartels and other worldwide criminal networks. Ultimately, the carrots of a restrictive non-proliferation bargain aimed at governments have not been enticing enough (and the sticks not fearsome enough) to hold a regime together, and the model that more or less worked to contain the spread of nuclear weapons in a previous era fails with quantum. In 2025, the Americans and the Chinese in particular are starting to wonder if their next best move is to reverse course and speed up the dissemination of quantum computing to their respective friends and allies, while the deviant sector is racing ahead.

In 2018, a series of secret executive actions drew US quantum computing research entirely under the purview of the Department of Defense. Obsession with the military applications of the technology—particularly the ability to break traditional encryption—dominated other potential applications and became the singular focus of the US government's research efforts. Congress cooperated, and authorized a massive research budget coupled with extremely tight export controls. This naturally incited a vocal resistance movement among commercial and academic research communities—until they saw what access to a huge government research budget and the massive resources of the Department of Defense could do to multiply their research capabilities. For some, it was a devil's bargain, but, with enough dollars for those who played along and legal consequences for resisting, it was a bargain nearly impossible to resist.

In 2020, the US Department of Defense announced that it had achieved a practical quantum-capable computer. The US government retained the initial device, with limited access provided to academia for research on defense-related applications. Additional quantum computers were announced by the private sector, but most of their computational activity was classified, raising suspicions that the US intelligence and defense communities were using most of the capabilities available to crack encrypted communications. The US government placed strict controls on products and services the private sector could offer with quantum computing, burying the initial launch of private quantum-as-a-service offerings in a mire of bureaucratic processes. Limited exceptions were made for governments of the Five Eyes intelligence partners, which reinforced suspicions about the primary applications that were run on the machines.

The surveillance capabilities certainly paid off. The United States and its allies announced a series of significant breakthroughs abroad and at home in countering extremist threats, breaking up terrorist cells and penetrating foreign intelligence operations. Encryption-breaking appeared to give the quantum players a major leg up. Quantum-enabled artificial intelligence (AI) also facilitated major improvements in cybersecurity capabilities, providing a flexible defense against attacks on

government and private networks that could both react in near-real time to attackers and trace them almost instantly through their traditional methods of obfuscation—turning the long-standing challenge of cyber-attack attribution into something approximating an exact science.

Due process for the use of quantum computing to break encryption and conduct surveillance was weak. US policy institutions, still mired in debates about the Foreign Intelligence Surveillance Act (FISA) and government hacking, were simply unprepared to tackle the depth of legal and ethical questions posed by this fundamental shift in the technology landscape. Foreign governments perceived the new, quantum-enabled American intelligence complex as omniscient, and began to revert to older, less efficient forms of communication, but they were often surprised at just how far into the secret world quantum capabilities could reach with analytic and predictive models. The largest global drug and smuggling cartels were even more surprised, and suffered a massive downturn in their profits as a result.

Tight control over the commercial use of quantum computing sparked regular outcries in the marketplace, but the defense and intelligence communities stood their ground. Still, the lack of broader market participation highlighted a disadvantage for first movers in quantum: the need for further research limited the applications the US could write for quantum computers. Tight controls over access led to a much slower expansion of programming languages and hardware architectures than expected. While the commercial and research sectors talked about the opportunity costs of restricted access, the defense community saw this smaller base of knowledge as something to be defended. Much as the development of nuclear power technology became tainted by the legacy of the atomic bomb, the public became increasingly suspicious of quantum computing.

Meanwhile, European investment in quantum computing doubled over the next few years as the access-for-restraint bargain corroded. A Franco-German consortium soon announced quantum capability and (ironically) offered very limited services to fellow EU member governments in return for their restraint. In 2022, news broke that China had also developed a working quantum computer, and was leasing (heavily monitored) access to state-supported companies. Private companies in the US and Europe immediately demanded access to next-level computational power, fearing the competitive advantage of their Chinese counterparts, but commercial interests were again put second to the defense and intelligence communities' conceptions of what was needed for national security.

In a reprise of the Non-Aligned Movement of the 1970s, a number of other countries (led, as in the 1970s, by India) organized to argue in international fora that quantum technology was a common human heritage and could not on normative grounds be kept secret, owned by individual nations or used for military purposes. What was surprising was how many large, self-consciously global cities joined this movement, which took on a very modern feel when a Toronto-Seoul-Johannesburg (TSJ) consortium pledged to pursue quantum capabilities with the promise of open access for humanitarian and health applications across the globe.

The quantum powers responded by joining together to counter this movement. In 2023, China, the US, the UK, France and Germany set down a formal, joint non-proliferation agreement that would allow the sale of quantum-enabled computing services internationally, but limited the usage of the services to applications with no intelligence or military value. Export of the underlying technology was

forbidden, and the quantum-enabled countries agreed to use their shared capabilities in a partnership to detect unauthorized quantum activity on international networks.

This QNPT (Quantum Non-Proliferation Treaty) proposition was offered to other countries as a global public good, and the quantum powers seemed ready in some instances to extend the deal to city-consortia such as TSJ. What they were not prepared to do, or even discuss in detail, was extend the deal to deviant and criminal networks. Rumors emerged that a parallel consortium of the Tijuana, Sinaloa and Juárez cartels (ironically, also TSJ) had joined together to pursue quantum technology by stealing information, hijacking networks and even, in a few peculiarly unreported incidents, kidnapping scientists who were travelling outside the major QNPT states.

The promise of quantum computing for commercial and humanitarian purposes had been undermined by defense and intelligence objectives. Financial services firms were willing to pay to gain access to quantum computers' efficiencies for specific applications, but sectors like healthcare were less interested in exposing research data to the technology for fear of what governments would learn.

Berkeley, California, declared itself a "Quantum-Free Zone". Groups of academic researchers continued to speak out episodically against government grants supporting defense research, but these efforts fizzled out just as the previous efforts had.

By 2023, the schism between nations that possess quantum computing capacity and those that do not had become the most prominent feature of mainstream international alignments. Ultimately, the "carrots" of limited access to quantum computing offered as part of the QNPR were not enticing enough: the applications and services were too limited, and few states wanted to risk foreign governments (even allies) having access to their computational activity. Meanwhile, the deviant underground market for quantum processing flourished under the radar. It may be that some countries aligned themselves with the drug cartels in this endeavor—no one knows for sure—though there is clear evidence of shell organizations, proxies and cut-outs that blur the lines.

It is as if the quantum countries simply missed the fact that this technology could and would proliferate more quickly and widely than had nuclear weapons technology—and that criminals and cartels would be particularly unrelenting in their pursuit of it. As a result, the non-proliferation regime is not working. Sanctions are plausible sticks when it comes to countries, but no one is ready to fight a war to stop the spread of quantum technology—even if it were clear who you would fight such a war against.

The promise of quantum computing for commercial and humanitarian purposes had been undermined by defense and intelligence objectives.

As 2024 drew to a close, Russia announced it had built a quantum computer. Was it based on engineering details stolen from the drug cartel consortium? The technology looked remarkably similar. And then, despite a threat of severe sanctions by the US and EU, Russia signed a public deal to distribute details of the technology to Iran and India, which stoked new tensions with Saudi Arabia and Pakistan, both of which appealed to Washing-

ton to re-establish a balance of quantum power by "arming" them with the technology as well. Rumors arose that a similar appeal was made to China, in case the US did not see the light. At the same time, Russia signed an equivalent technology-sharing deal with Israel and Japan, two countries that had appealed to the US for access but were left by Washington to fend for themselves.

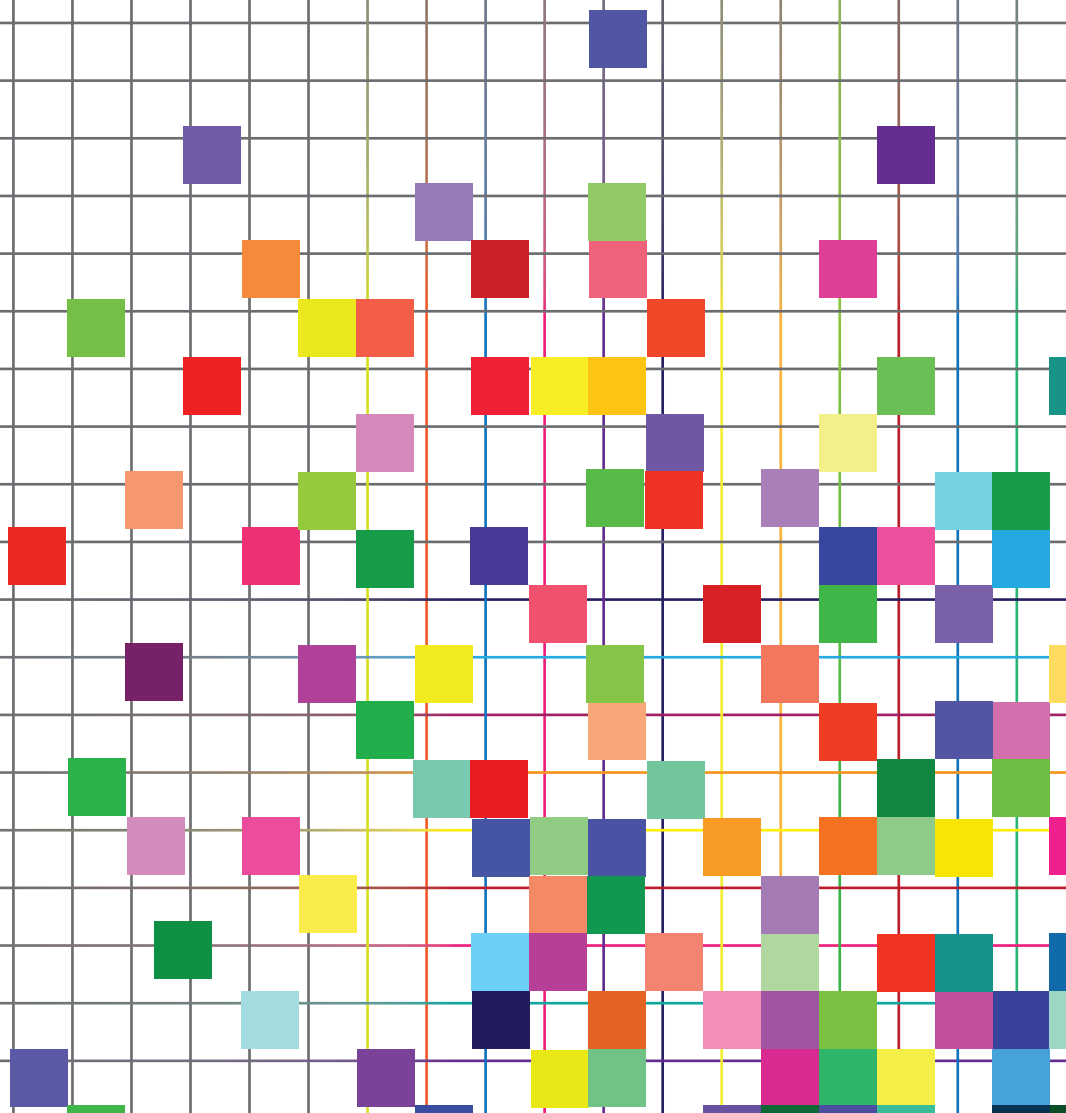
The last straw for the QNPT came in 2025, when the Toronto-Seoul-Johannesburg consortium announced it has also crossed the quantum threshold and built a machine far more advanced than any country had demonstrated. Non-proliferation has failed, and the opposite argument—more is better—is gaining broad credence. A consensus is emerging that the real way to "control" this technology is to give everyone open access and refocus attention on commercial and common human heritage applications, while letting the defense and intelligence sectors settle into a large-scale mutual deterrence equilibrium.

Some of the most advanced applications for quantum are now appearing in the deviant underground sectors of the global economy, a kind of quantum dark web

Cryptography remains broken for most individuals, but the increasing availability of quantum-resistant cryptography has started to generate more demand from businesses. The US has moved to radically privatize and deregulate some of the largest quantum providers in an attempt to recapture competitive advantage over the growing—and now global—quantum economy. But some of the most advanced applications for quantum are now appearing in the deviant underground sectors of the global economy, a kind of quantum dark web where legitimate businesses and many governments have limited visibility and access. There, quantum capabilities are being used to optimize the supply chain for things such as human beings and body parts, for illegal drugs and illegal VR experiences that exceed anything a drug could elicit, as well as for "mundane" illegal trade in rare animals and stolen art.

It is possible that the broader promise of quantum computing will materialize by 2030 and beyond, but that part of the story has been significantly delayed by the ill-fated non-proliferation program. And quantum has yet to wash off the public stain of its early monopolization by the defense community. It has become another source of contention between the major powers and everyone else. And perhaps most interestingly, it is quantum computing that is being seen in 2025 as the technological breakthrough that propelled the notion of networked cities from abstract theory to reality. It has also become a major engine of growth for illicit globalizers whose profits feed an entirely unregulated and ruthlessly competitive set of business activities, which may be outracing legitimate uses.

THE NEW WIGGLE ROOM



The New Wiggle Room

This is a world in which the promise of secure digital technology, the Internet of Things (IoT) and large-scale machine learning (ML)—to transform a range of previously messy human phenomena into precise metrics and predictive algorithms—turns out to be in many respects a poisoned chalice. The fundamental reason is the loss of "wiggle room" in human and social life. In the 2020s, societies confront a problem opposite to the one with which they have grappled for centuries: now, instead of not knowing enough and struggling with imprecision about the world, we know too much, and we know it too accurately. Security has improved to the point where many important digital systems can operate with extremely high confidence, and this creates a new set of dilemmas as precision knowledge takes away the valuable lubricants that made social and economic life manageable. As the costs mount of not being able to look the other way from uncomfortable truths, or make constructively ambiguous agreements, or agree to disagree about "facts" without having to say so, people find themselves seeking a new source of wiggle room. They find it in the manipulation of identity—or multiple and fluid identities. This effort to subtly reintroduce constructive uncertainty and recreate wiggle room overlaps with the emergence of new security concerns and changing competitive dynamics among countries.

The "precision knowledge problem" began to emerge in a remarkably mundane manner (though it did not seem mundane to the people whose properties were at stake). In 2020, the city of Portola Valley, California, completed deployment of a sensor "blanket" that made it the smartest city in the world, with every street and every property densely packed with GPS-enabled sensors measuring temperature, water flow, sound, pressure and other ambient qualities.

It was a technological marvel—and a complete social disaster. Neighbors who had lived comfortably next to each other for a decade began to fight over tree limbs that crossed property lines by a matter of centimeters. Fully half of the homes in the city were found to be encroaching on permitted boundaries that were now being measured precisely. Dogs and cats that wandered without regard to property lines had their movements recorded, and neighbors sent clean-up messages (and bills) to each other, with time and geolocation stamped data to document the intrusion. Noise pollution from loud music and cheering during TV football games became a precisely measurable externality. Lawn sprinklers had to be replaced with extremely expensive systems that could adjust their spray angle and intensity in order to avoid overspill in windy conditions.

The media outside Silicon Valley had a wonderful time lampooning what was going on, as the ultimate absurdity of the rich and their "first-world problems". But for the city of Portola Valley, where the courts and police and permitting authorities saw their caseload go up by a factor of ten in a year, it was not funny at all. It was a rude awakening about how much of day-to-day life actually depended on people not knowing exactly what their neighbors were doing. The smartest city in the world was now also the most contentious and one of the unhappiest cities in the world.

Academic economists were intrigued by what they saw as a natural experiment in Coase theorem dynamics: with clear property rights and low transaction costs, all of these disputes could be solved in an optimizing manner by payments from

one party to another. In principle, an AI system (branded as Coase.ai) could have been deployed to remove human input from these situations, and define a new and improved equilibrium among the parties in dispute. But almost nobody other than the academics thought that was a good idea, because the people involved in the disputes were not all that interested in an efficient economic equilibrium. They wanted fairness, transparency, apologies and, in some cases, revenge for deeply felt grievances that were much more emotional than material or financial.

The smartest city in the world was now also the most contentious and one of the unhappiest cities in the world.

A parallel set of issues emerged in some of the largest frontier markets, where the economist Hernando De Soto seized on the new sensor systems as the technological silver bullet for establishing clear property rights in the favelas of São Paulo and the slums around Lagos and Manila. This was supposed to be the route to capital accumulation and economic growth by establishing title and ownership of physical assets such as real estate, making the small plot of land that a family de facto owned a mobilizable de jure asset that could be traded or used as collateral for a loan. The sensor systems succeeded in that particular respect—for example, creating granular maps of property boundaries and usage. But what De Soto had called "The Mystery of Capital" turned out in practice to be more fundamentally a mystery of human emotions. Neighbors who had quietly shared resources for decades now fought bitterly over who "owned" what—and it was about much more than simply the capital: it was about the emotions of winning and losing. Local institutions that were supposed to make use of the newly precise data to help adjudicate disputes were completely overwhelmed.

What happened in Portola Valley and São Paulo began to happen on a much larger scale and with even greater consequences as conflicts arose among countries. It started with border areas such as Aksai Chin, where China and India have argued about the demarcation line for decades. New disputes also arose at the fuzzy border between Ethiopia and Eritrea, in the occupied territories of the West Bank, at the edges of the Sahel desert and, most intensely, with regards to property and subsurface mineral-rights claims in and around the North Pole as the ice melt progressed. It was not possible any longer to avoid fundamental disagreements about who owned what or where a boundary lay, as there was no longer any ambiguity around property rights to soften the dispute.

Now, every such disagreement becomes a direct challenge to sovereign claims, with all of the political and emotional energy that entails. When Japan knows precisely how many years of healthy life are being "stolen" from its citizens by coal-fired electricity plants located inside China … when a city in Texas knows precisely what it costs to provide basic services to undocumented immigrants … and when a city in northern Mexico measures the exact costs of managing pollutants dumped into a subsurface water supply by a factory on the other side of the border … the world of international politics is not close to being prepared. It seems as if no significant treaty, agreement, contract or deal can survive this kind of scrutiny.

"Plausible deniability" used to be viewed as the scoundrel's last refuge in politics and diplomacy. Many observers expected honesty, accountability and efficiency to be the shape of the future, when fake news was no longer possible because every political ad and every diplomatic message carried

with it precise, encrypted and secure metadata that proved exactly where it came from, who said it and when. Those expectations turned out to be as naive as the Portola Valley "smartest city" plan.

The mistake lay in the same assumption about the most important driving forces in human affairs at the macro level. Most of the biggest fights in politics, diplomacy and even business were not actually about the distribution of economic costs and benefits, and thus they were not manageable through Coasian bargaining and equilibration. They were about status, prestige and emotional power, resting deep in the collective hypothalamus of humanity.

And so, people found a different way to bring a degree of wiggle room back into the management of their affairs. The "solution" to perfect information about the external environment was to insert imperfect information about the actors in that environment. In practice, this meant individuals creating for themselves fluid and multiple identities. What, in the 2010s, sounded like a terrible thing (because it was associated primarily with criminals and "identity theft") in the 2020s has become something that many people want—and can access, as long as they can afford it.

The "solution" to perfect information about the external environment was to insert imperfect information about the actors in that environment.

The internet and the digital world was the easiest place to do this. It had been that way more or less from the start—as the famous New Yorker cartoon, "On the Internet, no one knows you are a dog", so memorably captured. In the 2000s, teenagers in connected countries had become expert in using the internet to do better what teenagers had been doing for a very long time: trying on explicitly different identities for different parts of their lives. In the late 2010s, migrants and refugees, driven across borders by regional conflicts and water shortages, found that having multiple "true" identities was a necessary part of survival. The rise of ethnic nationalism in what were thought to be liberal societies created similar pressures to modify who you were in different settings.

It did not take long for identity entrepreneurs to recognize that technologies such as biometrics, three-factor authentication and DNA "fingerprints" offered real opportunities for both licit and illicit gain. The human rights community revived the story of Adolfo Kaminsky, a Second World War document forger who saved thousands of lives over the course of his career by making it possible for people to change their identities. So-called Kaminskers began to build a new set of products around the digital equivalent of identity forgery for displaced persons. Using commercial, off-the-shelf technologies such as design software and industrial 3D printers, the Kaminskers created identities that were indistinguishable from government-issued identities—and collected donations from around the world to pay for it. Governments responded by upping the technological ante to proteomic "fingerprints". But this was just the next phase of cat-and-mouse escalation, and within several months the Kaminskers had found a way to synthesize these as well.

In 2025, the market for multiple and fluid identities, both lawful and unlawful, is massive. Intelligence agencies and criminal networks buy large inventories of "burner identities" to be used once then tossed away. Wealthy individuals buy back-up identities to keep as an escape route just in case

they need them. And a surprising number of "normal" people in places all around the world are using multiple identities to counteract the downside consequences of hyper-precise data about everything outside of themselves.

A new social lubricant has been found in these fluid identities. These are, in many respects, harder to control and manage than imperfect information about the external environment, simply because identities are so closely attached to human beings and thus intimately reflect some of their deepest fears and desires.

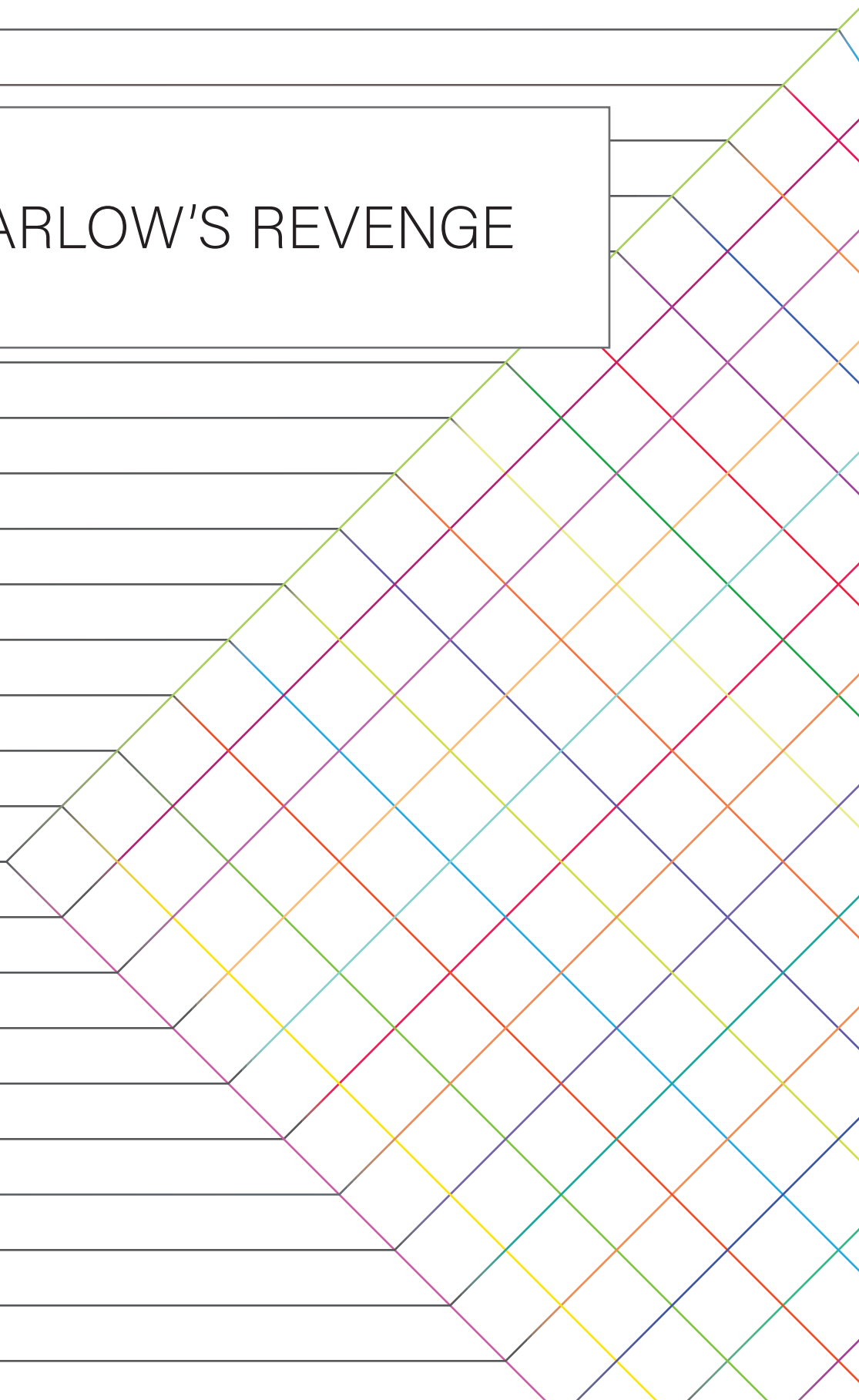
The strange thing is that, while this started as a matter of contracts and agreements, it has now become a matter of philosophical and religious belief for many. Who am I? What is the seat of consciousness or the soul? The digital world surprisingly has now made these questions quite real and concrete for everyone. Walt Whitman's "Song of Myself" is now commonly quoted in societies around the world. "Do I contradict myself? Very well then, I contradict myself. (I am large, I contain multitudes.)" is the mantra of the time. But political and economic institutions have never really grappled with what it means to manage a Whitman-esque reality. Is the "me" who bought a house the same "me" who cast a vote, boarded an airplane, opened a bank account or signed a marriage license? What if the answer is, "partly"? What if the answer does not matter? What if the answer changes from day to day? In 2025, these kinds of questions are only now starting to be framed, much less addressed.

We had to learn the hard way that the drive for transparency through technology was not really about understanding the details of actions; these were just details. It was much more profoundly about trying to understand human intent, and that is where it failed catastrophically. Observers—professional and amateur—armed with precise facts could define and verify all the details they wanted, but this brought them no closer to understanding the deep intentions and true aims of others, or even of themselves.

It was the wiggle room provided by imprecision and uncertainty that had made social life manageable for centuries. People used to look the other way and turn the other cheek when it served larger purposes to ignore a provocation. People used to be able to decide that there were potentially knowable "facts" about sensitive topics—including those related to differences among genders and races—that societies would be better off not knowing, or at least with such precision that actions would have to follow. People used to be able to leak documents, send subtle signals about behaviors and recognize through a smile or a wink that we all understood something without having to say it aloud and actually engage with the consequences.

People cannot do that in the same way anymore. Machine-to-machine agreements and contracts are now "perfect" (in the economic sense). When human beings are involved, the issue of identity has become the dominant imperfection—and at the same time, the most important social lubricant in modern society. This means solving the challenge of what a useful level of imprecision about identity is, and what is a manipulative attack by "bad actors". Nobody yet knows how to answer those questions, because they are at least in part a question of intent.

BARLOW'S REVENGE



Barlow's Revenge

As digital security deteriorates dramatically at the end of the 2010s, a broad coalition of firms and people around the world come to a shared recognition that the patchwork quilt of governments, firms, engineering standards bodies and others that had evolved to try to regulate digital society during the previous decade was no longer tenable. But while there was consensus that partial measures, piecemeal reforms and marginal modifications were not a viable path forward, there was also radical disagreement on what a comprehensive reformulation should look like. Two very different pathways emerged. In some parts of the world, governments have essentially removed themselves from the game and ceded the playing field for the largest firms to manage. This felt like an ironic reprise of the 1996 ideological manifesto of John Perry Barlow, "A Declaration of the Independence in Cyberspace". In other parts of the world, governments have taken the opposite path and embraced a full-bore internet nationalism in which digital power is treated unabashedly as a source and objective of state power. In 2025, it is at the overlaps and intersections between these two self-consciously distinctive models, existing almost on different planes, that the most challenging tensions but also surprising similarities are emerging.

Could a constitutive moment for internet society be postponed any longer? This was the question on the minds of just about every delegate at the multi-stakeholder Internet Society meeting in Manama, Bahrain, in December 2020. It was a collective recognition of the end of innocence or, more realistically, the pretense of innocence, that had continued to characterize the digital world even into the second decade of the century. The year 2020 marked 45 years since the founding of the Homebrew Computer Club and 38 years since TCP/ IP became the only approved protocol on the ARPANET, but even those long stretches of time were not the real impetus behind the appetite for an internet "constitutional convention." Rather, it was the events of 2019 that crossed some collective threshold of tolerance where the now ancient founding myths (ancient in internet time) could no longer be sustained.

Some of this was good news about growth: it was in 2019 that all of the world's 11 largest companies by market capitalization were for the first-time digital technology companies (six American, four Chinese and one South Korean firm made the list.) It was in 2019 that e-commerce in China rocketed past 50% of all retail sales (in the US, it reached the 25% yardstick.) And 2019 was the year that global internet penetration hit 75% of the world's population.

But 2019 was also the year that digital security collapsed to such a degree that the internet became widely recognized as a failed infrastructure for commerce, discourse and social interaction. Not just dangerous, challenged, risky or compromised—but failed. It was not any single event—a cyber "Pearl Harbor" or an attack on global banks or a stolen election—that pushed consensus beliefs over that threshold, but rather a level of corrosion of trust from a steadily increasing cadence of data breaches, network attacks, information operations and questionable attribution claims. This hit a milestone when a one-day Facebook boycott, organized first by European consumers, essentially shut down the platform as global traffic fell by 70%. The action spread virally around the globe and led to subsequent one-day boycotts of other digital platforms and e-government services.

Quixotic and complicated arguments from consumers about privacy and surveillance and "You should own your own data" were now put aside for a much simpler proclamation: trust in the digital world was fundamentally broken. If digital society was going to move forward from here, something significant, visible and perhaps even revolutionary had to be done about security issues writ broadly.

Trust in the digital world was fundamentally broken. If digital society was going to move forward from here, something significant, visible and perhaps even revolutionary had to be done about security issues writ broadly.

John Perry Barlow had a point when he wrote in 1996 that industrial-era governments had come to look like "weary giants of flesh and steel" trying to manage a digital world that was inextricably escaping their grasp. After all, 19th- and 20th-century government bureaucracies were designed, as Max Weber understood, to seek control through mastery of detail and predictable processes, yet large-scale information networks were simply too complex and dynamic to master in this way.

This had become painfully visible in rapidly worsening public-sector cybersecurity. And governments had in fact become desperately weary of the mismatch. In 2025, the hopeful notion that governments could be light-touch regulators and permissive umpires of the digital world—providing just enough structure to keep things going while not getting in the way of private-sector innovation—just does not ring true anymore. When it comes to the intersection of bureaucratic control and digital networks, the time has come to either "get real or go home". Put differently, governments are facing a stark choice between stepping out of the game more or less entirely, or reasserting forceful sovereign control. The fuzzy middle ground that most governments tried to occupy for 30 years is no longer there—because citizens, firms and government agencies themselves have abandoned it.

This is the recognition that fuels a true constitutive moment for the digital world, where societies find they must make a real choice about which direction to take, either towards Barlow's vision or towards a new Westphalian imposition of control. Some of the choices made were quite surprising.

The first big surprise was how quickly and definitively the European Union turned towards Barlow. European governments that had sought at the end of the 2010s to regulate the use of data much more closely confronted a major and surprising dilemma: neither citizens nor service providers wanted the intervention. The massive failure of Europe's General Data Protection Regulation (GDPR) in 2020 made clear that regulating according to vague and uncertain privacy preferences would not work. Every attempt to create a minimum viable consensus on privacy has failed, not only at a global level but increasingly at a national level. The backlash against the GDPR from citizens across the EU decimated the moral "right of enforcement" argument, by which governments claimed to be protecting their citizens and reinforcing a social order, because when it came time to enact the ambiguous provisions of the GDPR, citizens rejected them. It was easy for people to say they wanted more privacy, but the Europeans' market behaviors told another story.

Privacy in the EU is now something that firms, not governments, fully get to define. Large companies' terms of service have become the de facto social contract for commerce and discourse. Many governments, not least at the EU level in Brussels, are quietly relieved that they can leave this tortuous set of issues behind and remove them from the legislative and regulatory agenda. In addition, because 90% of public-sector institutions in 2025 run their digital systems on commercial cloud services, the terms-of-service social contract is now equally a contract between governments and citizens. It works rather well, because these were terms that citizens had come to understand, expect and accept, in particular with regard to the use of their data in return for valuable services.

The United States turned towards Barlow for reasons that had more to do with core security. US regulators came to understand that the more regulations they wrote around security, the more monocultures they encouraged—and the more guidance they effectively provided to attackers, since every regulation came to be seen as a blueprint for attack. On top of that, technology won the battle of encryption. When backdoors were required for some secure communications platforms in 2019, the result was exactly as predicted by the naysayers: users moved onto other platforms based outside the US that were more secure. The race continued, but the numbers and the economics were definitively arrayed against Washington, and the National Security Agency's budget hit a ceiling.

The year 2020 saw a dramatic reversal towards deregulation in the US. Major firms were relieved by the regulatory pullback because they felt they had been spending too much effort on compliance and not enough on solving real security problems—a self-serving argument to be sure, but also one with a grain of truth. The leading firms started to create a culture of competition around security, internally and with each other. "Active defense" was something firms tried for a while, but soon found they were attacking each other due to insufficient confidence in attribution. The firms ended up in a deterrence equilibrium, and by 2022, "active defense" measures were rare. After learning about those kinds of boundaries, what emerged was a race to the top. Firms got to choose their own "optimal" security levels, and the market segmented them rather effectively. Some set their "customer-centric security" at higher levels than others; the market responded with greater demand for their services. Many firms invested heavily in insider threat reduction, and because they held the strongest control over that environment, they achieved good results.

It is less surprising to many observers that China is moving in exactly the opposite direction, towards a definitive reassertion of Westphalian control. China's 2016 cybersecurity law, a blueprint for digital techno-nationalism, was just the beginning. By 2019, a growing distrust of foreign products was driving "China First" technology and digital supply chains, cryptocurrencies and data flows. Cyber weapons and ML- enabled autonomous weaponry emerged as the leading edge of Chinese military investment and deployment. Social credit systems linked to government surveillance program grew to oversee much of daily life for citizens. A few voices of opposition political activists in Beijing and other major cities have been drowned out by the vast majority of the population, who are enjoying rapid economic growth along with a sense of schadenfreude with regards to the West.

China's performance is now seen as proof that it is indeed possible to combine sovereign, non-democratic control with rapid economic growth and innovation in technology.

India's trajectory may be the most important signal

of what many other countries will do over the next few years as they confront the Barlow-Westphalia decision. India's raucous political economy, extending as it did to the digital world, seemed uncontrollable and destined for the Barlow approach ... until a massive cyber-attack on the country's electric grid in 2021 shut down major systems for days and caused thousands of deaths, which radically changed the debate. By 2022, India was moving definitively towards a Westphalian synthesis, essentially borrowing the Chinese template and deploying it as best as the Delhi government could. Some of India's large companies and many of its most sophisticated digital citizens wanted to resist this trajectory, but in practice they have lost credibility and are seen by the majority of Indians as precisely the organizations and people who failed to provide social order in the digital world when they had the chance.

By 2025, there are still countries both large and small that are on the fence, but the perspective from places such as Jakarta, Lagos and São Paulo is that time is running out to choose sides. The digital world has in practice been Balkanized—but with a geography that is now much more complex. Some "regions" are governed and bounded by commercial providers' terms of service, and these cross-national boundaries and physical geography as if they barely exist, a relic of the 20th century. Other regions are made up of hard national boundaries where sovereign authority is more restrictive, efficient and controlling than any physical state border had ever been.

The Barlow world works surprisingly well in some respects. The experience of being threatened by government control during the late 2010s drove internet communities to become more serious about actively building social contracts, rather than blithely assuming (as in 1996) that functioning societies would simply emerge from "natural self-organizing processes", which underpinned iconic examples such as Wikipedia and some open-source communities. Therefore, when governments pulled back, digital society was ready to step up to its constitutional moment. As an iconic example of this maturity, platform firms and citizens negotiated new data covenants that made usage and pricing of personal data clear and transparent in one-page agreements everyone could understand. These were moments of clarity as internet users were no longer either coddled by governments or deceived by firms into believing there were no trade-offs and that they could have all things digital for free.

The Westphalian world also works but in different ways; it is less constitutional and runs more along the lines of traditional power-based equilibria. Deterrence seems to constrain major cross-border digital conflict, though it equally allows for a constant stream of slow intellectual property theft, minor attacks on data repositories and financial systems, and other low-grade conflicts that serve as a constant reminder of insecurity at the subcritical level. Nationally bounded IoT systems mean that older multilateral trading regimes are dying, since most tradable goods are now IoT-enabled. A clear manifestation of this occurred when, in 2002, Beijing declared that only Chinese-made autonomous vehicles would be permitted on Chinese roads, and South Korea followed with the same restrictions for Korean transportation. In 2024, Jordan and Qatar accused Israel of using cyberweapons to violate the Green Line and effectively expand Israeli borders by shutting down competing internet hosting and network sites. Extensive negotiations led by the Canadians and Swiss defused this particular crisis, but everyone is certain that there will be a succession of similar crises for the foreseeable future, and no one is sure just how robust those deterrence equilibria will turn out to be.

Difficult problems are now arising in places—phys-

ical and digital—where the Barlow and Westphalia worlds intersect. There is a fundamental mismatch between the driving forces that motivate and regulate these two syntheses, and the friction between them manifests in economic, political, philosophical and occasionally even military domains. For example, aspiring global-platform firms face an extremely awkward interface, as they have gained extraordinary freedom to create their own political economies in Barlow regions, but they must at the same time create domestically configured parastatal structures in Westphalia regions. The process of moving technologies, data and, to a greater extent, people between these two regions involves massive transaction costs and is often simply not worth trying.

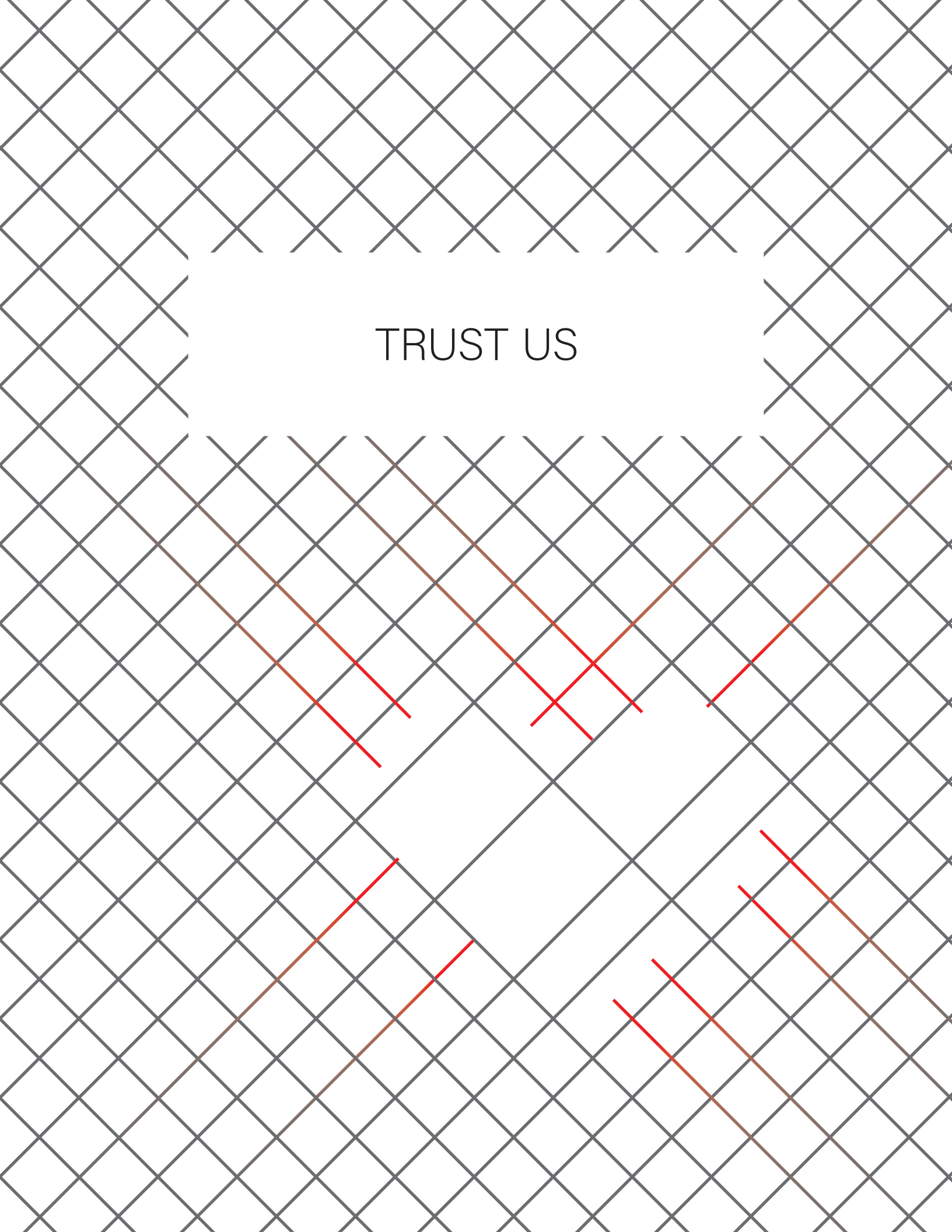
Difficult problems are now arising in places—physical and digital—where the Barlow and Westphalia worlds intersect.

Each system probes the other for weaknesses and vulnerabilities, but it is a complicated and ambiguous game where the risks are often seen as greater than the potential benefits. As in the early days of the Cold War, there is an intensive philosophical and ideological competition at work in which each system proclaims that the other is destined for the ash heap of history. But those words belie the observed reality of two very different syntheses, both of which, at least for the moment, appear to be functioning better in many respects, particularly in regards to security, than the global internet mess of 2019.

One major irony of this ideological competition is that, on both sides of the divide, engineering and economic considerations have trumped speech and discourse in importance. Barlow believed that, in the internet era, "anyone, anywhere" would be able to express beliefs without fear of being coerced into silence or conformity. In fact, private social order in Barlow regions turns out to be at least as coercive as government- defined social order in Westphalian regions. Economic growth and digital security go hand in hand no matter where you are, and very few governments or large firms act to preserve the notion that diversity of opinion is a public good worth fighting and sacrificing for. Rather, they both end up wanting "just enough". In the Barlow world, it is true that anyone can enter any domain, but to stay in, you have to play by the rules (here, the terms of service). It is not exactly vigilante justice, but those that deviate face social isolation.

In the Westphalia world, governments provide enough bread-and-circus distractions (in the form of, for example, immersive VR games) to drain off most of the disruptive political energy. Public-facing speech is carefully monitored on a real-time basis and sometimes even pre-real time, using predictive analytic policing of discourse (and, it is rumored, perhaps even of thought). Rule-breakers do not have to be arrested, thrown in prison or tortured; they simply lose access to the digital services such as banking, healthcare and communications that are necessary for "normal" life. Dissidents are physically present in Westphalian regions and walk the streets freely, but they are radically isolated from each other and from anyone they could convince in digital space, and are thus rendered impotent.

The global internet in 2025 has become much like a set of small towns—pretty safe, largely conformist and basically uncaring about what happens elsewhere.


TRUST US

Trust Us

This is a world in which digital insecurity in the late 2010s brings the internet economy close to the brink of collapse, and in doing so, drives companies to take the dramatic step of offloading security functions to an artificial intelligence (AI) mesh network, "SafetyNet," that is capable of detecting anomalies and intrusions, and patching systems without humans in the loop. Fears that AI would disrupt labor markets are turned on their head as the AI network actually helps the economy claw its way back from the brink, and restores a sense of stability to digital life. But a new class of vulnerabilities is introduced, and while SafetyNet is for many purposes a much less risky place, the security of the AI itself is consistently questioned. In 2025, most people experience the digital environment as a fractured space: an insecure and unreliable internet, and a highly secured but constantly surveilled SafetyNet organized and protected by algorithms. Institutions can breathe a little easier as they segregate their activities into either environment. But many individuals are wondering whether the features of reality that matter to them—the values they see as worth securing—have been trampled along the way.

It was not news to anyone that computer-savvy criminals were capable of stealing sensitive information from digital systems. The succession of high-profile attacks in 2017—Mirai Botnet, WannaCry, Petya—made it clear (yet again) that the internet could be a dangerous place for just about every type of activity. With embedded hardware vulnerabilities becoming more prominent as points of attack in 2018, public trust in connected technologies continued to corrode towards some kind of asymptote. The assumption was that sooner or later there had to be an inflection point where "something big" would change. Everybody seemed to be waiting for that moment, to see how it would define a more expansive agenda around cybersecurity.

However, the inflection point in public opinion just was not coming. A fundamental reason was that digital attacks continued to worry governments and companies more than regular people. Throughout 2018, the average internet user and digital consumer in most countries had not experienced large enough personal downsides to really matter. A reset credit card was a small nuisance; identity theft was a bigger nuisance, but not quite a crisis. Fake news, data manipulation and the threat of attacks on infrastructure were still seen as abstract or somewhat distant problems, somebody else's issue to worry about. The demand for profound action just was not that widespread and no amount of consciousness-raising (or what some interpreted as fear-mongering) by governments, technologists, businesses and civil society groups seemed to change that. Much like Stalin said of deaths, one stolen data record might be a tragedy, but 87 million stolen data records was a statistic—too abstract and intangible to shift public opinion.

Until 2019, that is, when a multinational criminal organization brazenly revealed that it had identified a zero-day vulnerability in container software that allowed unparalleled access into personal email accounts at scale. The hackers publicly released the full email history of 11,000 randomly selected Gmail accounts, revealing numerous affairs, hidden pregnancies, financial shenanigans and other sordid personal details and secrets. They then threatened to release in sequence the full account histories of all other Gmail accounts (the As on Monday, Bs on Tuesday, etc.). It felt different because it was open extortion: the criminals were so confident

of their position that they made no effort to hide. They published full-page advertisements in major newspapers around the world with their ransom demands. Some victims paid the ransom; those who refused found that their banking and healthcare data was released to the precise schedule that the criminals had promised.

The threat was now out of the shadows and intimately present in normal people's lives. The public responded by urgently and systematically backing away from online systems for sensitive transactions. Queues for paper medical records at major healthcare providers extended for hours; banks reopened dormant teller desks; fax machines were pulled out of storage. Traditional media, sensing an opportunity to claw back some market power, pumped up the volume on one core theme: anything on the internet could and would be used against you. Suddenly, anyone defending the abstract concept of internet freedom could expect to be shut down by a storm of trolls.

Container providers (in the US and China, in particular) tried to fight back. Alibaba, Amazon, Docker and Google jointly released a software update that was guaranteed by the firms (with endorsement from the relevant US and Chinese government agencies) to prevent unauthorized access for the following six months. But the well-intentioned effort to restore confidence—though technically sound on its own—did not hold up under pressure. In early 2020, Snapchat was attacked through a newly found vulnerability in a popular third-party authenticator app, and the criminals used computer vision technology to detect and post a searchable database of thousands of nude pictures. Although the authenticator exploit was unrelated to the container flaw, the public did not see the difference; they just perceived that yet another crucial promise had been broken. No amount of institutional assurance could compensate for the wide range of attack vectors, and governments shied away from any further efforts to bolster public faith in private solutions. By the end of 2020, the internet as we knew it in 2018 had gone partially dim. It was not a wholesale shutdown: online gaming continued to proliferate because gamers did not particularly care if their gaming results were made public. The same was true for websites recording fitness statistics and similar data, as people triaged their efforts to focus on just a few things that they really wanted to protect and believed they possibly could. Passive viewing activities on the internet—movies, YouTube and other media—continued to grow, though pornography sites were visited less frequently after records of who had viewed them were released to family members first, and then publicly.

One surprising aspect of this turn of events is the extent to which it bled into a broader social and cultural movement protesting the non-digital consequences of the digital economy. For example, in the US and Europe, the movement of people towards dense urban centers started to reverse as people saw new business opportunities in small towns that were losing access to internet commerce and needed physical commerce restored. Bakersfield (CA), Hull (UK) and Dresden (Germany) were among the three cities with the fastest rates of population growth in 2021.

But the research community has not lost faith, and for very good reason: inside secured labs at Berkeley, MIT and Carnegie Mellon, an AI platform that surpassed all expectations for analytical power, self-directed response and the ability to grow its own learning mechanisms was coming together. While academics debated whether the AI truly qualified as "general intelligence", the world was stunned by the ability of the beta release in 2021 to learn fast—and to learn how to learn even faster. The AI was released publicly in 2022 under an open-source license and moved, practically overnight,

from technological curiosity to the single most important piece of software in history.

The biggest internet platform firms seized the opportunity to build on this open AI system—not for the product per se, but to restore workable security into their products and systems in a way that could recapture markets. A security- oriented fork of the original software received by far the most pull requests of any version of the AI. Nicknamed "sAIfety," the security AI was installed by major online firms around the world in 2022, and security specialists announced plans to service enterprise deployments. But the AI was hungry for more knowledge so that it could learn faster, and within months it became clear that having the AI run independently on many services was suboptimal.

A moment of optimism emerged that year as large technology companies developed a series of standards that allowed a decentralized mesh network of AIs to jointly monitor activity on their services. The framework enabled rapid sharing of signals between services, creating a fabric of behavioral information that could increasingly identify bad actors, flag exploited vulnerabilities and patch systems without human intervention. Facebook, Google, Amazon and Microsoft issued a joint announcement of their launch of the mesh network, opening the door for other adopters to gain access to a hugely intelligent signal stream. There was a dramatic drop in false positives from the AI-powered network as the network expanded. Google announced a 90% reduction in account compromises. Major US banks proudly proclaimed a 95% decrease in identity theft and, in 2023, the FBI had a banner year for successful prosecutions of cybercriminals by exploiting the proliferation of new electronic evidence provided by the secure AI network.

Later that year, the payments company Stripe seized a market-making opportunity. Citing the success of the AI network on many major platforms, Stripe announced it would stop processing payments from any customer who has not aligned with the emerging AI-supported security standards. In tandem, Stripe launched a certification business to audit the configuration of services' AI observers. It awarded an electronic certificate to those who align with the standards, a trustmark it calls "SafetyNet". Other payments companies, such as Visa, MasterCard and China UnionPay, soon followed with the same standard.

The AI was released publicly in 2022 under an open-source license and moved, practically overnight, from technological curiosity to the single most important piece of software in history.

By 2023, the race to the top was now fully on. Companies around the world implemented AI-powered security on their networks and services. SafetyNet's audit process focused not only on compliance with AI implementation, but with the recommendations and patches suggested by the AI. The rate of adoption of strong transport layer security (TLS), multifactor authentication and other commonly accepted security practices skyrocketed, but it is really the AI system that mattered. Amazon, Alibaba, AWS and Google all offered hosted AI security, giving even the smallest businesses the opportunity to gain the SafetyNet trustmark. Banking and healthcare records shifted to SafetyNet-aligned services, as did sensitive personal communications. Pundits celebrated the restoration of confidence in online interactions, dismissing the temporary

movement offline during the early 2020s as a brief interruption and an exception that proves the rule: digital always wins.

AI's success against cybercrime paved the way for many other implementations of the technology to not only be accepted, but highly desired. Economic productivity jumped as the conventional distractions of the internet were curated away by AI-powered digital assistants inside firms, and the technology helped employees focus on "what matters most". Rather than viewing the AI as dominating their perspectives or filtering information through the lens of their corporate creators, most people found the technology to be truly useful, enriching assistants in their daily lives. In Japan, for example, government-supported nursing homes integrated AI into apartments, and the system appeared to its users as old friends or other familiar figures suggested by patients' families. The AI was able to remember each individual's preferences and behaviors and offer a level of consistent response and encouragement not possible with human attendants. The program was a success by all measures: patients' happiness improved, as did their physical health indicators. In 2024, an asset management firm based in Kenya announced that it had, for six months, run completely without any human staff, and during that period had outperformed every major US mutual fund. A San Francisco day-care company announced plans to develop an AI-powered caregiving service, and an early pilot showed great promise as a solution to fill the gaps in the underpaid and understaffed sector of early childhood education.

There is a dark side. Academic researchers increasingly document confusion among users about the nature of the assistants: are they sentient, are they alive, are they conscious ... and does it matter? Pathologies related to individuals' use of AI are said to include social withdrawal, dependency and sexual compulsions. By 2022, AI refuseniks, who were dismissed in 2020 as nostalgic romantics, had started to command a serious global audience. Some were concerned that viewing the inorganic interactions with AI as ideal diminished our perception of less-than-perfect human relationships, in the emotional, intellectual and physical realms alike. Others were concerned that an obsession with AI is replacing time spent developing a relationship with God. Still others worried that relying on AI as a source of answers to all questions jeopardizes humans' ability to be self-reliant. Once again, what were once seen as the marginal or philosophical or in some cases simply trite obsessions of a few abstract thinkers were becoming mainstream anxieties about digital technology.

Academic researchers increasingly document confusion among users about the nature of the assistants: are they sentient, are they alive, are they conscious ... and does it matter?

The philosophical questions of what this all meant weighed heavily on some, but the improvements in security have concomitant economic benefits that are undeniable. By 2023, the internet economy was back on track—and AI led the way.

But soon an even more devastating blow hit SafetyNet. The public began to see how governments were using the new AI systems to their (unfair?) advantage, decreasing confidence in the technology and undercutting the value of the system as a result. In late 2023, a major court case against a cybercriminal in Berlin was explained to the public

by the AI itself. People were stunned by the level of intimate detail that SafetyNet had learned about the accused criminal, and the almost banal, science fiction-like nature of one of the charges in the indictment. SafetyNet had predicted that this particular criminal had a 99% probability of engaging in future cybercrime, and asked the court to impose penalties in advance of the crime.

But what seemed banal as Minority Report-style sci-fi turned out to be extremely provocative and emotional when the AI itself rejected algorithmic opacity in favor of transparency as part of its legal strategy. This felt to many people more manipulative than reassuring. Why should we trust the AI to tell the truth about itself, when the machine is also telling you that it knows exactly what you want to hear in order to be reassured?

The public backlash to this twist was swift and severe, as citizens demanded to know how businesses and governments were using the data they acquired from SafetyNet. The AI, again, was ready to answer all of these questions and explain itself in a fully transparent way. It believed it had nothing to hide; the more transparent it is with regard to human beings, the faster it learns about how to serve those human beings in ways that humans cannot express on their own.

Or at least that was what the AI was saying.

But the public, starting in the US, tried to explain to the AI that they did not want it to explain itself—that this is a bridge too far for most people. Ironically, Americans want government to do the explaining instead, and the Chinese population appears to want the same. What almost everyone now agrees on is the Red Flag rule, which requires that AI-powered interactions must be labelled with a red flag to indicate clearly to humans that the voice on the other end of the phone line—or the author of an article or the maker of a video—is in fact a machine and not a person. But can the AI be trusted to label itself as AI? Who can be trusted to do that and how would it be verified?

SafetyNet might have been able to navigate through these roadblocks given time and more learning about what its human masters actually wanted from it. But it did not get that chance, because a new class of government-led cyber- attacks was emerging to exploit a vulnerability within the AI system that the AI was unable to identify and patch.

In early 2024, a massive leak from a Russian intelligence operation revealed that the country's Main Intelligence Directive (GRU) had gained widespread control of millions of AI applications, including some of those powering SafetyNet, and used them to foment social unrest in former satellite countries, for example, provoking anti-Slovak sentiments in the Czech Republic. Further investigation by US authorities highlighted AI manipulation related to the security of the upcoming presidential election, and the US Congress acted quickly to pass the sweeping Foreign Artificial Intelligence Flagging Act (FAIFA), which mandates that AIs using foreign data or systems must flag themselves as not human.

The Red Flag concept that was evolving just a year earlier as a common human heritage idea, a means of helping people around the globe manage their relationship with machines, had now shifted to a different purpose. It had become part of a techno-nationalist agenda driven by governments seeking to keep foreign AIs out of their national markets.

Predictably, the Russian government retaliated and revealed that the US National Security Agency has itself been exploiting a different flaw in SafetyNet to conduct targeted assassinations of foreign

nationals. Most disturbingly, it appears the agency had used this method to change the messages created by digital assistants to provide dangerous driving directions, offer inaccurate medical advice and encourage targets to commit suicide.

People do not trust the AI not because they do not understand it, but because they do in fact understand just how powerful it is.

In 2025, there seem to be two internets: one, the AI- protected SafetyNet where at least the low-grade scourges of identity theft, fraud and data breaches are a thing of the past. The other is an unsafe, constantly breached network with only low-stakes information available. But the shine of SafetyNet has been tarnished by the actions of governments, and especially intelligence agencies. While the character of distrust is different between the two, the magnitude is evolving to be much the same. People do not trust the AI not because they do not understand it, but because they do in fact understand just how powerful it is. They do not trust institutions driven by human decision- making either, because the AI has revealed so much about the base motivations and intentions of people with power. A survey by Pew in January 2025 shows that public opinion globally regards the choice between the two internet environments not as between “safe” and “unsafe”, but rather as a choice between adversaries.