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 labour 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 wasn't 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.

But the inflection point in public opinion just wasn't 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 wasn't 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 – didn't 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 wasn't a wholesale shutdown: online gaming continued to proliferate because gamers didn't 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 centres 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 hadn't 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 licence 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 behavioural 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.

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 behaviours and offer a level of consistent response and encouragement not possible with human attendants. The programme 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.

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 favour 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 can't 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 didn't 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 didn't 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.

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 don't trust the AI not because they don't understand it, but because they do in fact understand just how powerful it is. They don't 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.