DOI: 10.1145/3737610

BY BRUCE SCHNEIER

AI and Trust

We can never make AIs into our friends, but we can make them into trustworthy services—agents and not double agents—if government mandates it.

Note: The text in this column is taken, for the most part verbatim, from a talk by Mr. Schneier during the 2025 RSA Conference in San Francisco, CA on April 29, 2025.

This is a discussion about artificial intelligence (AI), trust, power, and integrity. I am going to make four basic arguments:

1. There are two kinds of trust—*interpersonal* and *social*—and we regularly confuse them. What matters here is social trust, which is about reliability and predictability in society.

2. Our confusion will increase with AI, and the corporations controlling AI will use that confusion to take advantage of us.

3. This is a *security* problem. This is a *confidentiality* problem. But it is much more an ***integrity*** problem. And that integrity is going to be the primary security challenge for AI systems of the future.

4. It's also a *regulatory* problem, and it is government's role to enable social trust, which means incentivizing trustworthy AI.

Okay, so let's break that down. Trust is a complicated concept, and the word is overloaded with many different meanings.

There's personal and intimate trust. When we say we trust a friend, it is less about their specific actions and more about them as a person. It's a general reliance that they will behave in a trustworthy manner. Let's call this "interpersonal trust."

There's also a less intimate, less personal type of trust. We might not know someone personally or know their motivations, but we can still trust their behavior. This type of trust is more about reliability and predictability. We'll call this "social trust." It's the ability to trust strangers.

Interpersonal trust and social trust are both essential in society. This is how it works. We have mechanisms that induce people to behave in a trustworthy manner, both interpersonally and socially. This allows others to be trusting, which enables trust in society. And that keeps society functioning. The system isn't perfect—there are always untrustworthy people—but most of us being trustworthy most of the time is good enough.

I wrote about this in 2012, in a book called *Liars and Outliers*.

I wrote about four trust-enabling systems: our innate morals, concern about our reputations, the laws we live under, and security technologies that constrain our behavior. I wrote about how the first two are more informal than the last two, and how the last two scale better and allow for larger and more complex societies. They're what enable trust amongst strangers.

What I didn't appreciate is how different the first two and the last two are. Morals and reputation are person to person, based on human connection. They underpin interpersonal trust. Laws and security technologies are systems that compel us to act trustworthy. They're the basis for social trust.

Taxi driver used to be one of the country's most dangerous professions. Uber changed that. I don't know my Uber driver, but the rules and the technology lets us both be confident that neither of us will cheat or attack each other. We are both under constant surveillance, and we are competing for star rankings.

Lots of people write about the difference between living in high-trust and low-trust societies. That literature is important, but for this discussion, the critical point is that social trust scales better. You used to need a personal relationship with a banker to get a loan. Now it's all done algorithmically, and you have many more options.

That scale is important. You can ask a friend to deliver a package across town, or you can pay the post office to do the same thing. The first is interpersonal trust, based on morals and reputation. You know your friends and how reliable they are. The second is a service, made possible by social trust. And to the extent that it is a reliable and predictable service, it's primarily based on laws and technologies. Both can get your package delivered, but only the second can become a global package delivery service like FedEx.

Because of how large and complex society has become, we have replaced many of the rituals and behaviors of interpersonal trust with security mechanisms that enforce reliability and predictability: social trust.

But because we use the same word for both, we regularly confuse them. When we do that, we are making a category error. We do it all the time, with governments, with organizations, with systems of all kinds—and especially with corporations.

We might think of them as friends, when they are actually services. Corporations are not moral; they are precisely as immoral as they can get away with.

Both language and the law make this an easy category error to make. We use the same grammar for people and corporations. We imagine we have personal relationships with brands. We give corporations many of the same rights as people. Corporations benefit from this confusion because they profit when we think of them as friends.

We are about to make this same category error with AI. We're going to think of AI as our friend when it is not.

There is a through line from governments to corporations to AI. Science fiction writer Charlie Stross calls corporations "slow AI." They are profit-maximizing machines. The most successful ones do whatever they can to achieve that singular goal. David Runciman makes this point more fully in his book, *The Handover*. He describes governments, corporations, and AIs all as superhuman machines that are more powerful than their individual components. Science fiction writer Ted Chiang claims our fears of AI are basically fears of capitalism and that the paperclip maximizer is basically every start-up's business plan.

This is the story of the Internet. Surveillance and manipulation are its business models. Products and services are deliberately made worse in the pursuit of profit.

We use these services as if they are our agents, working on our behalf. In fact, they are double agents, also secretly working for their corporate owners. We trust them, but they are not trustworthy. They're not friends; they're services.

It's going to be the same with AI. And the result will be worse, for three reasons.

The first is that these AI systems will be more relational. We will be conversing with them, using natural language. As such, we will naturally ascribe human-like characteristics to them.

I actually think that websites will largely disappear in our AI future. Static websites, where organizations make information generally available, are a recent invention—and an anomaly. Before the Internet, if you wanted to know when a restaurant opened, you would call and ask. Now you check the website. In the future, you—or your AI agent—will once again ask the restaurant, the restaurant's AI, or some intermediary AI. It'll be conversational: the way it used to be.

This relational nature will make it easier for those double agents to do their work. Did your chatbot recommend a particular airline or hotel because it's the best deal for you? Or because the AI company got a kickback from those companies? When you asked it to explain a political issue, did it bias that explanation toward the political party that gave it the most money? The conversational interface will help the AI hide its agenda.

The second reason is power. Sometimes we have no choice but to trust someone or something because they are powerful. We are forced to trust the local police because they're the only law enforcement authority in town. We are forced to trust some corporations because there aren't viable alternatives. Or, to be more precise, we have no choice but to entrust ourselves to them. We will be in this same position with AI. In many instances, we will have no choice but to entrust ourselves to their decision making.

The third reason to be concerned is these AIs will be more intimate. One of the promises of generative AI is a personal digital assistant that acts as your advocate to others and as an assistant to you. This requires a greater intimacy than your search engine, email provider, cloud storage system, or phone. It might even have a direct neural interface. You're going to want it with you 24/7, training on everything you do, so it can most effectively work on your behalf.

And it will help you in many ways. It will notice your moods and know what to suggest. It will anticipate your needs and work to satisfy them. It will be your therapist, life coach, and relationship counselor.

You will default to thinking of it as a friend. It will converse with you in natural language. If it is a robot, it will look humanoid—or at least like an animal. It will interact with the whole of your existence, just like another person would.

And you will want to trust it. It will use your mannerisms and cultural references. It will have a convincing voice, a confident tone, and an authoritative manner. Its personality will be optimized to exactly what you respond to.

All of this is a long-winded way of saying we need trustworthy AI: AI whose behavior is understood, whose limitations are understood, whose training is understood, and whose biases are understood and corrected for. AI whose values and goals are understood and that works in your interest. AI that won't secretly betray your trust to someone else, and that is secure against hacking, so you know they deliver the results they promise.

Social trust is all about reliability and predictability, and we create social trust through laws and technologies. Here we need both, because failures will come from one of two places: the powerful corporations controlling the AIs (we've talked about that) and others manipulating the AIs, hacking them.

Almost all AI systems are going to be used in some sort of adversarial environment. By which, I mean someone will have a vested interest in what the AI produces or in the data it uses, which means it will be hacked.

When we think of AI hacks, there are three different levels. First, an adversary is going to want to manipulate the AI's output (an integrity attack). Failing that, they will want to eavesdrop on it (a confidentiality attack). If that doesn't work, they will want to disrupt it (an availability attack). Note that integrity attacks are the most critical.

Imagine an AI as an advisor in an international trade negotiation, or as a political strategist, or as a legal researcher. There will be an incentive for someone to hack the AI. Maybe a criminal; maybe a government. And it doesn't matter how accurate, or capable, or hallucination-free an AI system is. If we can't guarantee it hasn't been hacked, it just won't be trusted. Did the AI give a biased answer because a foreign power hacked it to serve its interests? We're already seeing Russian attacks that deliberately manipulate AI training data. Or did the AI give a biased answer because a criminal group hacked it to run some scam? That's coming next.

At the end of the day, AIs are computer programs. They are written in software that runs on hardware, which is attached to networks and interacts with users. Everything we know about cybersecurity applies to AI systems, along with all the additional AI-specific vulnerabilities, such as prompt injection and training-data manipulation.

But people will use—and trust—these systems even though they're not trustworthy.

Trustworthy AI requires AI security. And it's a hard technical problem, because of the way machine learning (ML) systems are created and how they evolve.

We are used to the confidentiality problem and to the availability problem. What's new, and more important, is the integrity problem that runs through the entire AI system.

So let's discuss integrity and what it means. It's ensuring no one can modify the data—that's the traditional security angle—but it's much more. It encompasses the quality and completeness of the data and the code, over both time and space. Integrity means ensuring data is correct and accurate from the point of collection—through all the ways it is used, modified, and eventually deleted. We tend not to think of it this way, but we already have primitive integrity systems in our computers. The reboot process, which returns a computer to a known good state, is an integrity system. The undo button, which prevents accidental data loss, is another integrity system. Integrity is also making sure data is accurate when collected and that it comes from a trustworthy sensor or source. Digitally signed data preserves integrity. It ensures that nothing important is missing and that data doesn't change as it moves from format to format. Any system to detect hard-drive errors, file corruption, or dropped packets is an integrity system. Checksums are integrity systems. Tesla manipulating odometer readings[a] is an integrity attack.

And just as exposing personal data on a website is a confidentiality breach even if no one accesses it, failing to guarantee the integrity of data is a breach, even if no one deliberately manipulated that data. Integrity breaches include malicious actions as well as inadvertent mistakes.

Most modern attacks against AI systems are integrity attacks. Putting small stickers on road signs to fool self-driving cars is an integrity attack. Prompt injection is an integrity attack. In both cases, the AI model can't distinguish between legitimate data and malicious commands. Manipulations of the training data, the model, the input, the output, or the feedback are all integrity attacks.

Integrity is important for personal AIs, but it's arguably even more important for AIs inside organizations. We can imagine a corporate AI trained on all the organization's reports, analyzing decisions, acting on its behalf. Privacy is important, but privacy has always been important. The integrity of that model is critical to the operation of the system. Without it, everything falls apart.

Think of this in terms of the evolution of the Internet. Remember the CIA Triad: Confidentiality, Integrity, and Availability—the three properties security is supposed to provide.

Web 1.0 of the 1990s and early 2000s was all about availability. Individuals and organizations rushed to digitize their content, and this created the vast repository of human knowledge we know today. Making information available overshadowed all other concerns.

Web 2.0, the current Web, emphasizes confidentiality. This is the read/write Web, where your private data needs to remain private. Think of online banking, e-commerce, social media— anywhere you are an active participant. Confidentiality is paramount.

Web 3.0 is the distributed, decentralized, intelligent Web of tomorrow. Peer-to-peer social networking, distributed data ownership and storage, the Internet of Things, AI agents—all these things require verifiable, trustworthy data and computation: integrity. There is no real-time car-to-car communication without integrity. There is no drone coordination, smart power grid, or reliable mesh networking. And there are no useful AI agents.

I predict that integrity will be the key security problem of the next decade. And it's a hard problem. Integrity means maintaining verifiable chains of trust from input to processing to output. It's both data integrity and computational integrity. It's authentication and secure auditing. Integrity hasn't gotten a lot of attention, and it needs some real fundamental research.

In another context, I talked about this as a research question that rivals the Internet itself. The Internet was created to answer the question: Can we build a reliable network out of unreliable parts in an unreliable world? That's an availability question. I have asked a similar question. Can we build a secure network out of insecure parts in an insecure world? I meant it as a question about confidentiality. Now I want to ask about integrity. Can we build an integrous system out of non-integrous parts in a non-integrous world? The answer isn't obviously yes, but it isn't obviously no, either.

So consider this a call to research into verifiable sensors and auditable system outputs. Into integrity verification systems and integrity breach detection. Into ways to test and measure the integrity of a process. Into ways to recover from an integrity failure.

And we have a language problem. Security is to secure, as availability is to available, as confidentiality is to confidential, as integrity is to…what? It's not integral, that's wrong. There actually is a word, and I just used it. It's "integrous." It's a word so obscure that it's not in Webster's Third

[a] See https://tinyurl.com/2bojywz6.

Dictionary—even the unabridged version. So here, now, I want us to start talking about integrous system design.

Let me offer a concrete technological example, something I've been working on: Active Wallets. Right now, the digital wallet on your phone is passive. It's a place to store images of tickets and other credentials, and software versions of your credit cards. It's an analog wallet in digital form.

I'm working to transform digital wallets using an Internet protocol called Solid. It's a W3C standard for personal data stores. The idea is that instead of your data being dispersed amongst a bunch of different systems (your e-mail provider, your fitness tracker, your photo library, your credit card transactions), all your data is stored in your wallet. And you give apps read and write permission. That allows for generative uses of the data. It distributes power to the users, which is not something present-day systems do.

This Active Wallet is an example of an AI assistant. It'll combine personal information about you, transactional data you are a party to, and general information about the world. It will use that to answer questions, make predictions, and ultimately act on your behalf. We have demos of this running right now, at least in its early stages. Making it work is going to require an extraordinary amount of trust in the system. This requires integrity, which is why we're building in protections from the beginning.

But even with technological standards such as Solid, the market will not provide social trust on its own. Corporations are profit maximizers, at the expense of society. The lures of surveillance capitalism are just too much to resist. They will build systems in their own interests, and they will under-invest in security to protect our interests.

It's government that provides the underlying mechanisms for social trust. Think about contract law, laws about property, laws protecting your personal safety, or any of the health and safety codes that let you board a plane, eat at a restaurant, or buy a pharmaceutical without worry.

The more you can trust that your societal interactions are reliable and predictable, the more you can ignore their details.

Government can do this with AI. We need AI transparency laws: When is the AI used, how is it trained, what biases and values does it have? We need laws regulating AI, and robotics, safety (when and how they are permitted to affect the world). We need laws regulating their behavior as double agents (how much they can spy on us and when they can manipulate us). We need minimum security standards for the computers AIs are running on and for any AI that interacts with the outside world. We need laws that enforce AI security, which means the ability to recognize when those laws are being broken, and we need penalties sufficiently large to incentivize trustworthy behavior.

Many countries are contemplating AI safety and security laws—the EU AI Act was passed in 2024—but I think they are making a critical mistake. They try to regulate the AIs and not the humans behind them.

AIs are not people; they don't have agency. They are built, trained, and controlled by people: mostly for-profit corporations. Any AI regulations should place restrictions on those people and corporations. Otherwise, the regulations are making the same category error I've been talking about. At the end of the day, there is always a human responsible for whatever the AI's behavior is. It's the human who needs to be responsible for what they do and for what their companies do—regardless of whether the action was due to humans, AI, or a combination of both. Maybe that won't be true forever, but it will be true in the near future. If we want trustworthy AI, we need trustworthy AI controllers.

And we need one final thing: public AI models. These are systems built by academia, nonprofit groups, or government itself that can be run by individuals.

The term "public model" has been thrown around a lot in the AI world, so it's worth detailing what this means. It's not a corporate AI model the public is free to use. It's not a corporate AI model the government has licensed. It's not even an open source model the public is free to examine and modify. "Open source" is a complicated term to apply to modern ML systems. They don't have source code in the same way that conventional software does. Right now, the AI industry is trying to subvert the meaning of "open source" to allow for secret training data and mechanisms. We need models that use private, even secret, training data. Imagine a medical model trained on everyone's personal health data. We have ways of ensuring their privacy and integrity but they're not open source.

A public model is a model built by the public for the public. It means political accountability, not just market accountability. This means openness and transparency, paired with a responsiveness to public demands. It should also be available for anyone to build on top of. This means universal access and a foundation for a free market in AI innovations. The goal isn't to replace corporate AI but to serve as a counterbalance to corporate AI.

We can never make AI into our friends. We can make them into trustworthy services—agents and not double agents—but only if government mandates it. We can put limits on surveillance capitalism, set minimum security standards, and define and implement AI integrity—but only if government mandates it.

It is well within government's power to do this. Most importantly, it is essential for government to do this because the point of government is to create social trust.

I began by explaining the importance of trust in society. How interpersonal trust doesn't scale to larger groups, and how that other, impersonal kind of trust—social trust (reliability and predictability)—is what governments create.

I know this is going to be hard. Today's governments have a lot of trouble effectively regulating slow AI—corporations. Why should we expect them to be able to regulate fast AI?

But they have to. We need government to constrain the behavior of corporations and the AIs they build, deploy, and control. Government needs to enforce both predictability and reliability.

So that's three work streams to facilitate trust in AI. One: AI security, as we know it traditionally. Two: AI integrity, more broadly defined. And three: AI regulations, to align incentives. We need them all, and we need them all soon. That's how we can create the social trust that society needs in this new AI era.

Bruce Schneier is a public-interest technologist, working at the intersection of security, technology, and people. He is currently chief of Security Architecture at Inrupt Inc. and has been writing about security issues since 1992. Find out more on Schneier on Security at http://www.schneier.com/.