Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 706 7329
www.microsoft.com

Microsoft

Learning from other Domains to Advance AI Evaluation and Testing

We are grateful to the author of the enclosed expert report, which forms part of a broader series commissioned by Microsoft.

These reports were commissioned as part of Microsoft's effort to draw lessons from other domains to strengthen testing and evaluation as a cornerstone of AI governance.

The insights contained in each report reflect the authors' independent analysis and expertise. The views expressed are those of the authors alone.

We thank the authors for their intellectual generosity and thoughtful engagement throughout this project.

Full Series

Civil aviation: *Testing in Aircraft Design and Manufacturing* by Paul Alp.

Cybersecurity: *Cybersecurity Standards and Testing—Lessons for AI Safety and Security*, by Stewart Baker.

Financial services (bank stress testing): *The Evolving Use of Bank Stress Tests*, by Kathryn Judge.

Genome editing: *Governance of Genome Editing in Human Therapeutics and Agricultural Applications*, by Alta Charo and Andy Greenfield.

Medical devices: *Medical Device Testing: Regulatory Requirements, Evolution and Lessons for AI Governance*, by Mateo Aboy and Timo Minssen.

Nanoscience: *The regulatory landscape of nanoscience and nanotechnology, and applications to future AI regulation*, by Jennifer Dionne.

Nuclear energy: *Testing in the Nuclear Industry*, by Pablo Cantero and Gerónimo Poletto Antonacci.

Pharmaceuticals: *The History and Evolution of Testing in Pharmaceutical Regulation*, by Daniel Benamouzig and Daniel Carpenter.

March 13, 2025

Cybersecurity Standards and Testing -- Lessons for AI Safety and Security

By
Stewart Baker
stewart@stewartbaker.net
+1 (202) 709-6402

Introduction

There was little need to worry about the security of computer systems until the 1960s. Before that, computers were hulking machines locked in a room that only a few trusted boffins could enter. That all changed when time-sharing debuted, allowing multiple users to use the computer at more or less the same time. That posed the risk that they'd start looking over each other's shoulders. And that led defense and intelligence customers to wonder how they could protect their classified data from ordinary users.

Sixty years on, we are still trying to answer that question.

Many experts were sure that the answer was to set security standards and enforce them by testing systems to see whether they met those standards. That is still the closest thing we have to an answer, but it hasn't been a very good one; it's at best a partial success. The story of its failures is in some ways the story of politics and policy writ large at the turn of the twenty-first century; as such, it may also tell us a lot about how AI safety and security standards will succeed, and fail.

A theory of technology regulation

New technologies all end up being regulated in some way. What kind of regulation is a function of the era in which the technology first comes to wide attention.

To take one example, regulation of radio and other broadcast media could have been limited to a few rules about avoiding signal interference; instead, the broadcast regulatory regime adopted in the 1920s and 30s incorporate the full scope of Progressive politics, including the now-familiar fear that nationwide radio and TV companies would displace mainstream sources of news and narrative like local papers. In other examples, automobile regulation was alive from the first to the dangers of crime and anonymity on the roads, but it took sixty years to address driver safety. And the earliest regulations for railroad and telegraph companies embraced a "public utility" model because in the Gilded Age what their users most feared was discriminatory or exclusionary prices. Indeed, as telephones became common, utility regulation had such mindshare that the entire industry was more or less forced into becoming a monopoly for the convenience of regulators who were comfortable with that kind of regulation.

The effort to regulate computer security is a child of very different times. In the 1960s, when time sharing on Big Iron machines began to create security concerns, the computer industry was still being driven by government customers (though that would soon change). In that context, agencies like ARPA, NSA and other classified agencies believed they could solve the problem simply by designing and buying a secure computer system on their own. That first effort focused on to developing a Multics system with a secure kernel. It failed, largely because the burgeoning commercial market for computers reduced enthusiasm for a bespoke "government computer" design. In an effort to accommodate the evolution of a private market, security experts tried to turn their Multics design into a set of computer security standards that vendors could meet for government and other security-conscious customers. These were ultimately memorialized in the NSA's 1983 "Orange Book," which provided for testing to evaluate a computer system's level of security. NSA – History of Computer Security, declassified manuscript (11 February 1998).

By the 1980s, though, the commercial market for computing had begun to dwarf the government market, and hostility to government surveillance had grown markedly. The times had truly changed. After the Vietnam war, Watergate, and post-Watergate intelligence investigations, NSA was no longer automatically trusted to define computer security. Indeed, for some computer scientists and hackers, cybersecurity meant deploying unbreakable encryption so NSA couldn't read network traffic.

Skepticism of government peaked in the neoliberal 1990s. Deregulation of industry that began under Carter and Reagan proved to be a success. The idea of imposing regulation on a dynamic and exciting new industry like computers was a nonstarter. If anything, industry argued, the key to security was deregulation -- an end to restrictions on encryption exports, which NSA had used to assist its global surveillance mission.

Industry won that fight, but cybersecurity problem just got worse. The George W. Bush, Obama, and Trump administrations all searched for ways to fix the problem without regulation, even as more – and more serious – intrusions on computer networks came to light. What they ultimately all settled upon was remarkably similar -- "light touch" security standards that relied heavily on industry to fill in the gaps and even grade its own homework, with only a threat of government sanctions if industry failed to take its responsibilities seriously.

Because efforts to avoid AI security and safety problems are likely to encounter the same environment, the middling successes of the regime for cybersecurity regulation are likely to provide valuable lessons for government as it deals with artificial intelligence.

Government procurement standards

NSA took the lead in setting security standards for computer systems, ultimately creating the "Orange Book," which specified a range of possible security levels with letter grades from D to A. The intent was for industry to build to the standards and then have their products tested for compliance.

That was easier said than done. Divisions soon arose over how strictly to specify and interpret the security criteria. The parties that were charged with evaluating product compliance needed more detailed documentation than anyone had expected; otherwise, they could not really understand the systems they were judging. Correspondence and paperwork proliferated, delays accumulated, and the results were uninspiring. Often, by the time a system was fully evaluated and deemed secure under the Orange Book, it was commercially obsolete. The original government expectation was that vendors would quickly achieve a grade of C and then move to the harder work of meeting the criteria for the B level. In fact, getting a B proved dauntingly difficult and time-consuming, and the commercial market provided only a modest boost for those who did. NSA had to lower its expectations, hoping it could get vendors to at least a verified high C.

Even that watered-down security impact attracted international criticism and ultimately even more dilution of the original vision. US allies objected to the fact that testing was only open to US computer vendors. At their insistence, the Orange Book was replaced by "Common Criteria" – security standards that had to be negotiated among Western governments. Compliance with these homogenized standards were then judged by a range of evaluating bodies, with each government seeking to use its own home-grown body of evaluators.

In the end, it was unclear how much extra security was being provided by the Common Criteria's heavy infrastructure. Soon, the evaluators more or less abandoned the idea of holding everyone to a single clear standard; instead, by the 2010s they had moved to a system in which vendors offered relatively narrow security promises and then the evaluators determined whether those promises had been met.

NIST Cybersecurity Framework

While the Common Criteria were narrowing their focus, concern about security had expanded beyond hardware and software to encompass entire organizations and their networks. By 2014, attacks on government and private company networks were everywhere. Institutions proved incapable of preventing the extraction and misuse of data they had collected about individuals. To incentivize better protection for user data, the government again looked to security standards.

This time the standards were written by the National Institute of Standards and Technology, or NIST. Its Cybersecurity Framework reflects lessons drawn from the problems suffered by the Common Criteria. Unfortunately, one of the lessons was that much discretion about the evaluation must be left to the party being judged.

The Framework is as layered as an onion. The outer layer identifies at a high level the kinds of activities that a security-conscious organization must conduct – things like identifying its networked assets, protecting them, detecting attacks, responding to attacks, recovering from attacks, and putting in place a governance structure to oversee those activities. Steven B. Lipner, The Birth and Death of the Orange Book, IEEE Annals of the History of Computing (April-June 2015). Each layer is broken down into further layers -- categories and subcategories of activity. Organizations being evaluated must specify which categories and subcategories they wish to be

judged on. They must also specify how strictly they want their security in those categories to be graded. As a result, achieving "compliance" with the Cybersecurity Framework is a highly flexible notion. Nonetheless, there are a number of evaluators involved in judging whether an organization meets the requirements of the Framework. Internal compliance teams conduct self-assessments. External auditors also conduct compliance reviews. And regulators may monitor compliance as well.

But because of the flexibility built into the specification process, the Framework may better be understood as offering institutions a way to think comprehensively about improving their security posture, rather than setting enforceable security requirements. As a result, while compliance with the Framework may be evaluated by auditors, they often are closer to coaches or advisers than to teachers issuing pass-fail grades. For its detractors, then, the Framework and its auditors are not a substitute for demanding regulation or performance standards. Defenders of the Framework in contrast would say that tough regulations and high-performance standards are a pipedream, and all that can be expected is a tool for nudging organizations toward good-faith efforts at better security.

Regulating Private Sector Security

Hostility to regulation remains a significant force in shaping governments' approaches to cybersecurity. Cybersecurity simply was not a serious concern during the New Deal and Great Society eras when much regulatory legislation was adopted. And adding security to existing regulatory requirements means passing new legislation, not an easy lift in today's Congress. As a result, regulators have to base their security rules on general and sometimes questionable legislative authority. This has pushed agencies away from detailed requirements and demanding audits.

Financial institutions: The Gramm–Leach–Bliley Act (GLBA). The 1999 Financial Services Modernization Act included general provisions requiring that financial institutions protect the privacy and security of customer data. These general provisions formed the basis of a security rule issued by the FTC and enforced by several financial regulatory agencies. The rule provides guidance on how to develop and implement specific aspects of an overall information security program, such as access controls, authentication, and encryption. It also imposes institutional governance requirements, such as a security officer, a comprehensive security program based on foreseeable risks, an incident response plan and annual security reports to the board of directors.

Financial institutions are required to conduct regular risk assessments and to test and monitor the effectiveness of the security program. Congressional Research Service, Banking, Data Privacy, and Cybersecurity Regulation, February 24, 2023 They must monitor and maintain logs of user activity on information systems. They must also conduct regular vulnerability assessments or penetration tests of their systems. That said, the actual security measures that must be in place are not spelled out. Given the wide range of capabilities in the financial sector, the rule calls only for a security plan with "administrative, technical, and physical safeguards that are appropriate to

your size and complexity, the nature and scope of your activities, and the sensitivity of any customer information at issue."

The FTC has formal enforcement authority for the rule but its resources for testing and evaluating compliance are insufficient to police the sector's cybersecurity. Practical testing of institutions' security and provision of detailed guidance depends heavily on the role of bank examiners who work for financial regulators such as the Federal Reserve. Examiners conduct on-site reviews of how the FTC's general guidance is being interpreted. Because examiners probe the practices of multiple institutions, they often act as "best practices" missionaries with a stick.

Healthcare: The Health Insurance Portability and Accountability Act (HIPAA). The Health Insurance Portability and Accountability Act of 1996 (HIPAA) authorized the U.S. Department of Health and Human Services (HHS) to develop regulations protecting the privacy and security of certain health information. The result was the HIPAA Security Rule, later amended in response to the HITECH Act.

The HHS Security Rule leans toward formal, prescriptive rulemaking, specifying things like the use of encryption to protect health information, the contents of contracts with business associates, and the detailed and obnoxious attestations that require patients to agree to the disclosure of their health information. Even so, most of the actual cybersecurity measures identified in the Security Rule are not so specific. Rather, the rule identifies broad areas of security concern, such as workforce security, contingency planning, access controls, access control, audits, and policies. Under these headings, relatively general guidance is provided, and further details must be supplied by the regulated institutions themselves.

That said, the Security Rule features aggressive and detailed compliance testing of a kind familiar to administrative lawyers. HHS emphasizes cooperation and offers assistance to regulated entities. But it also encourages anyone who identifies a violation to file complaints with the agency. In addition to investigating complaints, HHS also conducts compliance reviews. Regulated entities must keep records and submit reports as necessary to determine compliance, and must cooperate in any investigation. Enforcement procedures are formal and quasijudicial. HHS allows entities to make formal submissions of evidence and of defenses. Parties may request a hearing before an administrative law judge; they have the right of discovery and cross-examination of witnesses. HHS may serve subpoenas to obtain testimony and documents as needed. Testimony is transcribed. At the end of the process, civil money penalties may be imposed for violations.

Bulk Power Generation: FERC CIP standards. In 2008, the Federal Energy Regulatory Commission adopted "Critical Infrastructure Protection" (CIP) cybersecurity standards. Mandatory Reliability Standards for Critical Infrastructure Protection, Order No. 706, 122 FERC ¶ 61,040, The standards were originally developed by industry consensus through the nonprofit North American Electric Reliability Corporation, or NERC There are a total of thirteen FERC standards identifying industry best practice in these areas:

- The elements of bulk power cyber systems that must be identified

- The kinds of security controls that must be adopted
- Topics for security management training
- Rules for electronic security perimeters
- Rules for physical security of bulk power premises
- Requirements for managing security systems (ports, patches, malware, alerts)
- Incident reporting, response planning
- Recovery plans
- Configuration change management
- Security of information about grid systems
- Security of communications between control centers
- Supply chain risk management
- Physical security risk assessments

While the thirteen standards contain some specific requirements, much guidance is necessarily general or consists of examples of possible compliance measures. Detailed enforcement of the requirements is the province of the North American Electric Reliability Corporation, or NERC. NERC and its regional entities audit, investigate, and assess operator compliance. Based on the severity of the violation, NERC can impose sanctions, order immediate corrective actions and mitigation measures. The Federal Energy Regulatory Commission, or FERC, plays a classic regulatory role, including conducting audits of power suppliers. These audits include interviews with operators and compliance staff as well as field visits. Auditors regularly identify individual or industry-wide security weaknesses. These result in recommendations for changes in practice or security standards.

Pipelines: Transportation Security Administration (TSA). TSA has authority to oversee the security of oil and gas pipelines, which exercises through a regularly updated security directive. Despite industry unhappiness with some of the early requirements, TSA's authority over industry cybersecurity is on stronger ground than many other agencies'; the need for better pipeline cybersecurity was also dramatized by the serious economic and political fallout from a ransomware attack on Colonial Pipeline in 2021. As a result, TSA's guidelines have a clear political and legal foundation, and this has enabled the agency to promulgate relatively specific and testable security requirements.

Even so, the most recent directive, Security Directive Pipeline-2021-02E, issued on July 26, 2024, does not list all security measures and technologies that must be adopted and that will form the basis of later TSA inspections to determine compliance with the plan. Instead, as with earlier directives, the guidelines call for a relatively detailed cybersecurity implementation *plan* that will be reviewed and approved by TSA.

Scope of the plan. Preparation of the plan requires that the operator of a pipeline identify critical systems, external connections, and zone boundaries for both IT and operational technology systems. The operator must have an access control plan for enforcing security controls on boundaries between zones. The plan must include identification and authentication policies, multifactor authentication, least privilege and separation of duties, and limits on shared accounts,

along with a schedule for reviewing existing domain trust relationships. The operator also must adopt continuous monitoring and detection policies that include a variety of capabilities, with particular emphasis on auditing and logging suspect network activities and an ability to isolate industrial controls systems that pose a cybersecurity risk. Finally, the plan must address the need for prioritization and application of security patches and a fallback plan for mitigating security holes that cannot be patched without degrading business critical functions.

Other plans. Additional plans are also required. One is a cybersecurity incident response plan, with clearly designated roles and regular exercises to ensure that they are understood. These exercises must test a significant part of the plan each year. Companies must also have a plan for regularly assessing and auditing their cybersecurity. This assessment includes an architecture design review, verification of network traffic, and a system log analysis. It also calls for penetration testing that includes adversarial red-teaming. The assessments must cover at least a third of all the policies in the operator's cybersecurity implementation plan. The results of the assessment must be included in an annual report to TSA.

TSA has the authority to test the operator's performance by inspecting its records. To eliminate doubt, particular documents that TSA may inspect are listed. They include log files, a capture of network traffic for 24 hours or less, and snapshots of traffic within operational technology systems and between those systems and information technology systems.

Lessons for AI Safety and Security Regulation

The cybersecurity story has many parallels in the effort to establish safety and security standards for artificial intelligence. Cybersecurity failures have persuaded many that industry must do better, and governments have tried many ways to incentivize improvement. But a regulation-skeptical zeitgeist and a complex, rapidly changing technical environment make it impossible for regulators to set hard and fast network security rules – or for them to be too aggressive in policing the rules they do adopt. Cybersecurity regulators do have an option that may not be as easily available to AI safety and security regulators: It's clear that cybersecurity failures will turn up as regularly as Metro buses; so, regulators can simply wait for the next security disaster and harshly punish the company deemed responsible. True AI safety and security disasters may not lend themselves to such an approach.

More likely, AI regulators will be pushed toward the same approaches as cybersecurity regulators. Among these, the following features of cybersecurity regulation seem likely models for AI regulation and testing:

- A broadly cooperative and consensus-driven approach to setting safety standards.
- Reliance on the "best practices" of competitors to give content to standards.
- Requiring companies to create their own comprehensive safety and security plans in accord with standards that are mostly abstract concepts rather than concrete requirements (and that can be evaded or diluted by companies that do not apply them earnestly).
    - This doesn't rule out the possibility that, in an ocean of vagueness, the standards may contain islands of specificity, much like the cybersecurity regulators'

requirements of encryption or multi-factor authentication. These specific provisions will be more easily tested and thus perhaps somewhat overemphasized in implementation.

- Requiring a corporate security officer and team that is empowered to police compliance and held accountable for initial enforcement and testing of the company's implementation of safety and security standards.
- Requiring that records of compliance failures be kept for inspection and that the safety and security officer make regular reports to the board of directors on corporate compliance with the safety and security standards and plan
- Encouragement and certification of a third-party consulting industry that tests compliance and provides support to internal security compliance teams.
- Accommodation of other nations' interest in advancing their own national champions by diluting the standards and the testing to accommodate those champions.
- A stick in the closet: a regulatory body with heavy (if undefined) authority to impose penalties for noncompliance, including failure to conduct appropriate testing.

Third-Party Security Assessments

Even "light-touch" cybersecurity regulation recognizes that compliance ultimately depends on external oversight. However cooperative and flexible the regulatory regime may be, companies are rarely left to grade their own homework. But regulatory resources were scarce in the Biden administration, and even scarcer as the Trump administration imposes deep staff cuts. This, plus a desire to avoid bureaucratic inflexibility, has led to more interest in having third parties assess compliance with cybersecurity standards. Not every regulatory regime has embraced third-party assessment, however, and those that did are discovering that introducing third parties introduces new complications.

The traditional approach – government assessment. As our earlier tour of representative cybersecurity regulations showed, most regulatory agencies reserve to themselves the authority to assess compliance. The one exception identified above is TSA, which requires penetration testing and adversarial red-teaming – a specialized skill that neither the agency nor most companies is likely to possess.

Toe in the water. Some regulatory regimes have embraced at least the possibility of third-party assessment without fully committing.

The European Union and UK have adopted a Network and Information Security (NIS2) Directive that is intended to set cybersecurity regulatory standards for all EU members. While intended to set an achievable baseline for security, NIS2 has proven controversial as small and medium businesses are pulled into its orbit. (Indeed, Germany's governing coalition was unable to pass a law implementing the directive, and it now faces infringement actions for missing the deadline for implementation.)

NIS2 expects state regulators to supervise companies' cybersecurity practices, but it also contemplates that regulators may conduct supervision by requiring "targeted security audits

carried out by an independent body or a competent authority." NIS2, art. 33(2)(b). And it expressly authorizes agencies to require that the regulated company cover the costs of such audits. Such cost-shifting from the regulator to the regulated is no doubt one of the attractions of third-party assessment. And article 33 can be read as encouraging experimentation without mandating such assessments.

Mandatory third-party assessment. The Defense Department, with massive numbers of defense contractors to oversee, has gone farthest in delegating assessments.

FedRAMP. The Federal Risk and Authorization Management Program (FedRAMP) sets security standards for cloud-based services. It makes third-party assessment organizations (3PAOs) central to its regime. Certification of third-party assessors is demanding, and the stakes for cloud providers are high.

The initial impetus for FedRAMP establishment in 2011, was to replace siloed agency evaluations with a one-time authorization process that all agencies could trust and reuse, setting a common security baseline for cloud-based services while also reducing redundant efforts. By 2012 the program had reached initial operating capacity, meaning the program had the processes and staff in place to start granting authorizations. Since then, the program has undergone numerous changes to improve its efficiency and scale. By 2016 approvals were taking over a year and a FedRAMP Accelerated initiative was launched with the goal of cutting these in half. This was followed by further efforts to streamline low-risk approvals (lowering the barrier to entry for small-medium enterprises) and to use more automation in the approval process. Despite these and other improvements, FedRAMP is still widely criticized in 2025 for being too manual, time-consuming, and expensive, leading to the launch of another initiative, FedRAMP 20x, a pilot program to streamline assessment and compliance.

FedRAMP is costly for both federal agencies and their contractors. In a GAO study of the program, one cloud provider reported spending $367 thousand on its third-party assessor. Another estimated that it spent $3 million on overall compliance, including the cost of its assessor. In the same study, cloud providers complained that despite their cost, assessors were not always fully conversant with the security standards and sometimes adopted inconsistent interpretations of FedRAMP requirements, differences that could greatly increase costs.

Despite (or perhaps because of) these challenges, FedRAMP may provide the best model for how a mandatory third-party security assessment system would operate.

CMMC. The Cybersecurity Maturity Model Certification (CMMC) is the U.S. Defense Department's program to enforce unified cybersecurity standards across its contractors. It was introduced in 2020 to ensure companies handling sensitive unclassified defense information meet specific security benchmarks. Unlike prior self-attestations, CMMC requires independent third-party assessments for most contractors as a condition of contract award, meaning companies must earn a certification at the appropriate level or risk being ineligible for DoD contracts. The stakes are therefore high for defense suppliers to comply.

CMMC 1.0 rolled out in early 2020 with an ambitious five-tier model of cybersecurity maturity, but it quickly proved too complex and burdensome for many, especially smaller subcontractors. The program's goal was to replace patchwork self-certifications with a single, trustable standard ("certify once, use many") for the entire defense supply chain. However, by 2021 the rollout had stalled – only a few pilot assessments occurred, and no contracts yet enforced CMMC – as industry pushback and slow accreditation of assessors prompted the Pentagon to pause for an internal review. In November 2021, after considering public feedback, DoD unveiled CMMC 2.0, a streamlined version designed to improve scalability and adoption. CMMC 2.0 condensed the model from 5 levels to 3, eliminated unique CMMC-specific process requirements (fully aligning with existing NIST SP 800-171/172 controls), and introduced more flexibility in enforcement. Notably, under 2.0 even Level 1 (basic safeguarding) can be achieved with annual self-assessments affirmed by company leadership instead of requiring outside auditors, and many mid-level contractors handling less critical data will also be allowed to self-certify rather than undergo third-party assessments. These changes were intended to roughly halve the compliance burden for companies while still protecting sensitive data, making it easier for small and mid-sized businesses to participate in defense contracts without sacrificing cybersecurity. CMMC 2.0's revisions were codified through rulemaking in late 2024, and DoD set a phased timeline so that by 2025–2026 most new DoD solicitations will include CMMC requirements tied to contract eligibility.

Implementing CMMC can be expensive and challenging, especially for smaller firms. For example, DoD's own analysis estimated that achieving a mid-tier certification (the level needed to handle controlled unclassified information) would cost over $118,000 in the first year for a typical contractor. In a 2021 industry survey, 24% of U.S. electronics manufacturers said the costs and burdens of CMMC might force them out of the defense market entirely. More than half of the surveyed suppliers indicated that any compliance cost above $100,000 would be unaffordable – a threshold below the DoD's projected cost – and about one-third expected it would take one to two years of preparation before they were ready for a CMMC audit. Early on, companies also reported confusion and inconsistent guidance in the CMMC 1.0 process, as different assessors or consultants sometimes had varying interpretations of requirements, adding to anxiety. The DoD has acknowledged these concerns; under CMMC 2.0 it moved to provide clearer guidance, allow Plans of Action for minor shortfalls, and offer support resources to help businesses meet the standards. These steps aim to reduce the risk of smaller contractors being priced out of the defense industrial base due to cybersecurity mandates.

Non-regulatory requirements for third-party security assessments. Regulators are not the only parties imposing security requirements. Buyers of goods and services are discovering that they are only as secure as their least secure supplier. This has led large buyers to demand security assurances and to explore ways of enforcing those assurances. Not surprisingly, as revealed in an RSA Executive Security Action Forum study, these have ended up in the same place as security regulators, calling first for information through questionnaires and self-assessments and then for more objective assessments from third parties. And all of the criticisms aimed at regulatory assessments have been hurled at these private regimes: they are too expensive, too time-consuming, inconsistent from one buyer to the next, and too focused on checklists rather than security results.

Lessons for AI regulation. The popularity of third-party assessments in cybersecurity means that the same mechanism will be attractive to governments seeking to regulate AI without the disadvantages of traditional regulation – great expense, a reduction in competition, regulatory capture and misuse of regulatory standards, and the like. To be blunt, there is no free lunch here. Cybersecurity regulation demonstrates that using third-party assessors has some advantages. It certainly costs less -- from the government's point of view. Some of the costs are shifted to the regulated party, but competition among assessors is likely to reduce the total cost. That would seem to be a gain, but another risk of using third-party assessors is that they will compete not just to lower the price of an assessment but also to lower the stringency of the assessment.

Other issues for AI regulation are raised by experience with third-party cybersecurity assessments. Complaints about ambiguities in security standards, for example, will be even more pointed in the context of AI safety and security, topics that have attracted a multitude of competing aspirations. It is impossible to use third-party AI security assessors if the standards for AI security have not been defined in detail. This author, at least, has seen no sign that AI security and safety have been defined with specificity.