FEBRUARY 2022

Data Protection or Data Utility?

Cryptographic Software Solutions for U.S. Innovation Competitiveness

By Alexander Kersten and Isaac A. Robinson

Despite the economic upheaval wrought by the Covid-19 pandemic, the United States is now poised to enter a new period of productivity growth with the widespread deployment of artificial intelligence (AI) technologies. And just as past eras' technologies have been driven entirely by the consumption of resources such as steam, coal, oil, and natural gas, this AI-driven period will rely on data. Yet, while this data-driven economic growth model has proven benefits to both individuals and firms, it also raises serious concerns over individuals' data privacy.

A key challenge in this regard is to ensure that data is handled securely and that the privacy of American citizens is protected. Because of the regard paid to data protection and privacy over the past decade, the proliferation of digitization has made privacy policy a part of innovation policy. However, until now, policymakers have viewed data use and data protection as trade-offs, with some nations adopting strict control of data flows. These measures have not been effective in practice; strict curbs have stifled innovation while doing relatively little to protect privacy.

An alternative is to deploy promising cryptographic software solutions that can enhance privacy while still allowing access to data. This solution not only unlocks the commercial potential of data for use by firms, law enforcement, nonprofits, and researchers but also protects individuals' privacy. Realizing the benefits of this win-win solution will require further development and widespread adoption of emerging cryptographic software solutions.

> *And just as past eras' technologies have been driven entirely by the consumption of resources such as steam, coal, oil, and natural gas, this AI-driven period will rely on data.*

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

Productivity Gains from Big Data

Widespread deployment of AI and machine learning (ML) technologies can contribute to gains in U.S. productivity. In fact, the **combination** of AI and "Big Data" alone is expected to lead to the automation of nearly 80 percent of all physical work, 70 percent of data processing, and 64 percent of data collection tasks.

As these technologies mature, AI and user data appear set to follow a well-documented "**J-curve**" growth. This trajectory sees general purpose technologies undergo initial periods of tepid growth associated with early investment and adjustments while integrating the new technology, followed by rapid integration and a dramatic increase in productivity. The past 150 years of U.S. economic history have witnessed similar periods of total factor productivity (TFP) **growth** following the widespread adoption of new general-purpose technologies such as electricity, internal combustion engines, and then with information and communications **technologies** in the 1990s.

The Debate on Trade-offs

Given that these emerging technologies exploit data, a central question for this next phase of potential growth centers on concerns related to the privacy and protection of information. In general, data security has two parallel aspects: privacy and protection. Privacy is a person's ability to determine what information is collected about them and how that information is used. Protection is how well the data is secured and how its uses align with user preferences once it has already been **gathered**.

This trade-off between data security and data utility is being tackled in various ways across the globe. In 2018, the European Union rolled out its General Data Protection Regulation (**GDPR**) with the goal of giving Europeans better control over their data. Meanwhile, on November 1, 2021, China took major steps to **implement** its new Personal Information Protection Law (**PIPL**), which looks similar to GDPR but with greater government control of cross-border data flows. In the United States, these data concerns were highlighted in President Joe Biden's July 2021 **executive order** "Promoting Competition in the American Economy." In particular, the executive order calls for ending "unfair data collection and surveillance practices that may damage competition, consumer autonomy, and consumer privacy." In addition, the United States Innovation and Competition Act (**USICA**), which passed in the Senate in June 2021, devotes Section 2673 to the "protection of data and information from public disclosure."

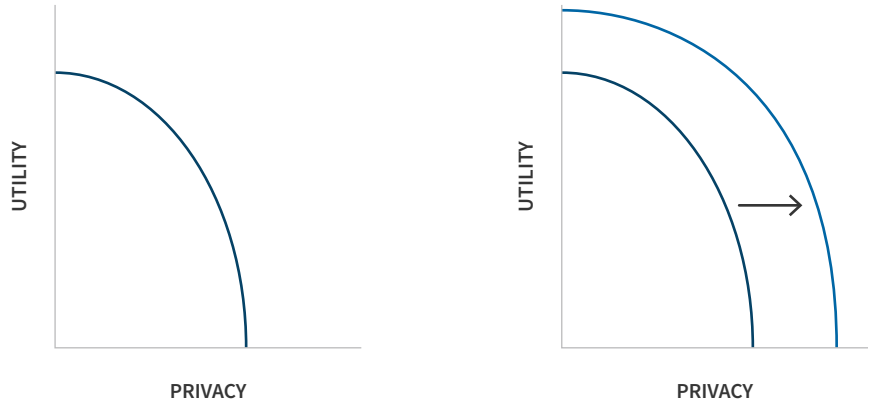
U.S. policies regarding data privacy and protection can benefit from what the United States can learn from others. Indeed, strict data privacy **regulations** have been shown to impose significant economic and social costs by burdening innovative small companies and start-ups. Regulations such as the European Union's GDPR have already shown to have adverse effects on **free speech**, consumer choice, and even **scientific research**. Further, **experts** have noted that the GDPR does little to protect privacy and instead focuses almost exclusively on its stated goal of data protection.

In considering a balanced policy approach, U.S. policymakers should consider the following realities. First, it is important to recognize that AI is becoming increasingly accessible as tools become easier to use and cloud-computing prices continue to fall. *Wired* magazine's founding executive editor and famous technologist Kevin Kelly **claims** that "the business plans of the next 10,000 start-ups are easy to forecast: *Take X and add AI*." The policy challenges created by AI need to be addressed.

> *Strict data privacy regulations have been shown to impose significant economic and social costs by burdening innovative small companies and start-ups.*

Second, several U.S. states and firms have begun to take the privacy issue into their own hands. California notably moved forward in 2018 with the California Consumer Privacy Act (CCPA), which gives Californians the right to delete personal data, opt-out of the sale of personal data, and to know how businesses use their data. In January 2021, Apple announced new privacy changes to its IDFA (Identity for Advertisers), which gives consumers control over whether or not they want to share their personal data. In August 2019, Google announced Project Sandbox to develop open standards for enhancing privacy on the web, which has since been followed by Google's browser Chrome promising to phase out the use of third-party cookies by mid-2023. These cookies, which track user behavior across the web, are essential to digital marketers, who as of 2021 relied on Chrome as the most widely used browser to still allow first- and third-party cookies. On the other side, Facebook, mobile marketers, brands, and game developers contend that these privacy changes stifle their ability to do business through targeted advertising.

The Data Protection-Utility Curve



Source: Authors' creation.

Third, and perhaps most fundamentally, the standard approach to framing the issue may be outmoded. The trade-off is traditionally seen as one of maintaining privacy or allowing for the flourishing of technological innovation and subsequent economic growth—with one coming at the cost of the other. In this way, the issue has been framed on a Pareto boundary along the so-called "privacy-utility curve." Privacy regimes and data protection regulations in Europe and the United States have wrestled with this assumed trade-off, and the United States has struggled to decide where it should fall along this curve. The policy lens must instead be refocused on the question of how the United States can innovate to regain its competitive lead and national security advantage in a rapidly changing industry.

While some technology firms have taken aspects of data privacy into their own hands, U.S. policymakers have to date found it difficult to justify more stringent data privacy laws in light of the potential negative impacts on innovation and economic growth.

The Role of Advanced Software Solutions

Glaringly absent amid this debate, save for the most technical of audiences, are software solutions that can maximize data privacy while addressing important security concerns and minimizing negative impacts on innovation, economic development, or scientific research. In doing so, these solutions effectively shift the curve to the right, making their discussion and development just as high a priority.

Arati Prabhakar—a former director of the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST), and now a founder of the nonprofit Actuate—pointed out in a recorded conversation with CSIS president and CEO John Hamre that software innovations can be used as solutions to questions of data privacy. She noted that data sharing among different sectors in the United States is held up by mistrust and division between the government, the public, and private organizations, which makes the type of data collection needed for social innovation difficult. Yet, the ability to use data to improve access to education, healthcare, and other services is the key to the country's success as it grows larger and more diverse this century. This means that software's ability to anonymize and encrypt data is of particular importance.

Many of these software innovations require significant bandwidth and computational power, which in the past have rendered them practically infeasible. However, new advancements in hardware, such as those explored by DARPA's new DPRIVE (Data Protection in Virtual Environments) program, promise to make this possible in the coming years.

Dr. Prabhakar and her colleagues at Actuate think that the United States is on a track to more "democratized data" but also that the lack of an adequate political or legal framework for harnessing this data remains a hindrance to more widespread implementation. Policymakers, therefore, have a clear role to play in helping guide the United States in this century-defining move. Cryptographic software that can enhance privacy while still allowing access to data can not only unlock commercial potential but deliver vast public benefits by protecting individuals' privacy while allowing for valuable insights hidden in the vast stores of data that firms, law enforcement, nonprofits, and researchers have at their fingertips.

Software as a Data Privacy Solution

DATA PRIVACY

Data privacy options are used to properly handle sensitive and personal data. Relatively novel cryptographic software options offer solutions to the purported trade-off between greater consumer privacy and innovation while promising data protection and privacy for the sake of national security.

The idea of software as a privacy solution has two breakthrough areas: secure multiparty computation and consent management. For the United States' market-driven approach to data security, these developments are a welcome and promising sign.

- **Data Privacy Management Software:** Data privacy management software is a growing space where privacy management firms check compliance and collect consent. Yet, at the time of this writing, they are unable to handle the increased data flow facilitated by secure computation, since users would be inundated by consent forms. This often leads to either users having to sign away broad privacy rights or data not being used (one issue that many have brought up in regards to the GDPR). There is a large existing market for data privacy management software, including companies such as OneTrust and DataGrail.

- **Consent Software Agents:** Consent software agents are a suggested solution and active area of research that involves software agents that would automate the consent-granting process. Users set preferences, and the program uses an algorithm to determine whether to grant consent. This alleviates the flood-gate problem of needing to grant consent every time a person's data is transferred, used, or sold.

  Consent software agents would effectively create a virtual "data box" that sits in the homes of individuals. Personal data would be safeguarded within this data box, and whenever someone wanted to use it, they would ask the data box for permission. This would both allow the individual to determine where their data is being used *after* it is initially collected or after the original point of sale. It would also facilitate increased logging of where in the internet universe data is being stored and used, which could help facilitate new and existing laws, such as the right to be forgotten, which is part of the GDPR.

  Unfortunately, this solution also suffers from a lack of laws and policies designating culpability. Namely, it remains unclear under current American jurisprudence whether the producers of consent software agents would be liable for all consent decisions made by their products. Moreover, such a system would require significant digital infrastructure upgrades for data-using and data-sharing companies, investments they may be unwilling to make without appropriate legal motivation.

Software as a Data Protection Solution

Data can exist in three different states: at rest, in transit, and in use. While standard encryption is used for the first two, it is data in use that is most useful but also most vulnerable. This is because encryption creates a "ciphertext" that obscures data's plaintext, meaning that good encryption works only if it produces ciphertext that is seemingly completely uncorrelated to its plaintext, appearing as something useless or incomprehensible. This works with data at rest and in transit, but to perform a mathematical operation on the data (that is, to be able to "use" it), one either needs to decrypt the data or utilize some relationship that exists between the encrypted data and the plaintext. Current solutions almost universally decrypt the data to use it, but this exposes unencrypted and sometimes private data, leading to legal and ethical challenges as well as increasing vulnerability to nefarious actors.

PROTECTION FOR DATA STORAGE

Beyond the well-known solutions for protection for data storage (data at rest and in transit), such as the encryption and identity access management that most people are familiar with, there are now new approaches introduced by blockchain.

- **Encryption:** Encryption is a tool equivalent to "password protecting" data, and it works by creating a "ciphertext" that obscures the data's plaintext, effectively obscuring the data. This works well for data at rest and in transit but not in use.

- **Identity and Access Management:** Identity and access management systems regulate which users within an organization have access to data, whether over the cloud or on-site. Given the shortcomings of password protection, even given a strong username and password, most people are now familiar with using multifactor authentication (MFA), which is being converted into identity and access management products. Today, identity management systems often have elements of biometrics, AI and ML, and risk-based authentication.

- **Blockchain and Decentralization:** With its decentralized form of data storage, blockchain is almost impervious to hackers, though its current implementation of an immutable ledger does mean that

data cannot be removed or edited, not allowing the "right to forget" online that the GDPR enshrines. Blockchain has garnered particular interest around the use of ledgers for online payments using cryptocurrencies such as Bitcoin.

Protection for Data in Use

Technologies that allow data to remain encrypted while in use offer a real breakthrough in technical and social innovation since they do not jeopardize user privacy. In light of privacy laws in many places, access to certain data might be hard to obtain, or scientists and other data users may not be expert data managers or may have doubts about their own ability to prevent data leaks. But with the following options, especially fully homomorphic encryption (FHE), an analyst could do calculations on troves of data without having to ever reveal underlying data that might expose the identities or other personal information of individuals. Further, these methods have their basis in lattice-based cryptography, meaning that they are of the NP-hard level of computational complexity. While to humans, this simply means that they are intractably hard problems to solve—involving millions of points over tens of thousands of dimensions—they are also hard for quantum computers to solve and, therefore, can be expected to play an ongoing role as the basis of post-quantum computing cryptography.

Overcoming the issues of ciphertext, privacy, and even the lack of trust among institutions in the United States, secure multiparty computation and FHE are the best technical solutions to the field of privacy-preserving computation.

- **Secure Multiparty Computation:** Secure multiparty computation (SMC) is like blockchain in that it distributes computation along the network so that no one needs access to the full data set to get a result. It allows users to compute pieces of data across multiple machines, which is a way of protecting individual bits. This would allow users to keep their data in their own private space while at the same time contributing associated insights to the ecosystem. Therefore, a bad actor would have to collect everything to make sense of the original data, which would be practically impossible.

- **Fully Homomorphic Encryption:** Fully homomorphic encryption is a type of encryption that preserves the structure of the data, allowing the user to analyze and gain insights from the data without decrypting it and exposing the personal information—not unlike a locked glove box where one can manipulate what is inside but not access it. A solution for FHE was only just discovered in 2009 and is six orders of magnitude more computationally intensive than simply partial homomorphic encryption.

- **Federated Learning:** Anyone who has ever experienced keyboard word prediction improving through continued use has likely experienced the effects of federated learning. Federated learning is a technique for decentralized training of ML models that allows such models to be trained on private data without ever having to put that data online or in "the cloud." In federated learning, a centralized model is trained on some pre-attained or synthetic data. Then, a version of this model is sent out to a select group of online devices. These devices run and train the model locally on the device's private data. Once the model is trained, each of these devices sends their updated version of the model back to the centralized source. The distinction here is that these devices are sending the model, basically a long list of numbers that represent what the model has learned, rather than the data itself.

Potential Pitfalls and Current Limits of the Software Solution

Yet, many of these approaches, especially the most promising types such as multi-party computation (MPC) and FHE, are still computationally expensive and require a high level of expertise to implement

properly. Currently, these methods require those with PhDs to implement, but they are edging closer to commercial viability. FHE and MPC have intense mathematical rigor that makes them solid solutions to protecting data, but FHE is computationally heavy, which severely limits where and when it can be used. Meanwhile, MPC has a bandwidth-heavy problem because of its approach to spreading compute pieces across various computers. Given these limitations, researchers attempting to use these methods need to decide what the time requirement of the computation is and what the limitations of their resources are, whether computation or bandwidth. Therefore, as computer scientists such as those at DARPA's DPRIVE work toward making these truly viable solutions, analysts hoping to apply these cryptographic schemes should find ways to combine both.

But might there still be ways for bad actors to access these data? Unfortunately, looking at "**output privacy**," with infinite queries and perfect logic, immoral actors can reverse ML models and extract private information from them. The current solution is to add random "noise" to the data, which makes the model less accurate, thus making it harder for bad actors to extract information. Further, computation cannot be completely anonymized, as data scientists need to know the structure of the data to determine what models they will build and which tests they will run. To address this, the current solution is data generation, which uses secure computation to generate fake data that researchers can use to build their models. The models are then applied to the original data set without researchers having access to individual-level or sensitive data. However, the solution to these issues may be a "plug-and-play" approach to provide a way for untrusted people to use the data without having full access to the data. "**DataSafes**," as they are called, offer a complete system with plug-and-play secure computation, output privacy controls that prevent deanonymization, and data generation for preliminary model building. With this promising solution, each institution, including university researchers, can perform studies on all this data without accessing anyone else's data. Therefore, there is certainly a path forward where data mining and privacy can coexist—seemingly a contradiction to many today.

The Way Forward on Cryptographic Software

Policymakers should follow the technological developments summarized above to advance objectives in national security and international competitiveness. Indeed, a lack of agreement around the governance and protection of data privacy itself creates a national security vulnerability. There are many who **believe** a federal privacy law would provide overdue security to American citizens from foreign **threats**.

The 2021 final **report** of the National Security Commission on Artificial Intelligence found that "without adequate data protection, AI makes it harder for anyone to hide his or her financial situation, patterns of daily life, relationships, health, and even emotions," which can become national security weaknesses as well. With the world's population coming online in growing numbers, there now exist unprecedented troves of data on where people live, what they do, what they want, and almost everything else. In a world where hard borders have grown porous and the digital privacy of a nation's citizens can be threatened by bad actors anywhere in the world, national security calls for security in cyberspace as well.

Today, the United States can lead the next phase of innovation, built on the ethical use of data for research and development. Just as previous technological revolutions have been driven by various forms of energy—whether steam, coal, liquid natural gas, or renewables—so will this new breakthrough era of AI and information technology be driven by data and information.

Further development and promotion of these cryptographic software tools are the key to the United States maintaining its leadership in the innovation race while also taking steps to uphold data security.

Encouraging the more widespread adoption of new and existing cryptographic software, especially the game-changing, up-and-coming methods for data in use, such as FHE and SMC, is of major importance. U.S. policymakers should support organizations that are working on breakthroughs in effective and viable cryptographic software solutions that support innovation and protect privacy. ■

> *Further development and promotion of these cryptographic software tools are the key to the United States maintaining its leadership in the innovation race while also taking steps to uphold data security.*

Alexander Kersten *is deputy director of the Renewing American Innovation at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Isaac A. Robinson*** *is a former researcher with the CSIS Economics Program and is a current student at Harvard University.*

This brief is made possible by general support to CSIS. No direct sponsorship contributed to this brief.