Four Steps to Preserving Privacy and Debiasing Data-Informed Policy

Brian R. Spisak, Joseph Spisak, and Andrew Trask

Spisak, B.R., Spisak, J., & Trask, A. (2021). Four Steps to Preserving Privacy and Debiasing Data-Informed Policy. *California Management Review*. Available at https://cmr.berkeley.edu/2021/01/four-steps-to-preserving-privacy-and-debiasing-data-informed-policy/

Addressing Privacy and Bias with Computational Leadership Science (CLS)

California Management Review

TECHNOLOGY

Four Steps to Preserving Privacy and Debiasing Data-Informed Policy

by Brian R. Spisak, Joseph Spisak, and Andrew Trask

Machine learning can be a force for good, or a tool of surveillance and oppression.

☑ INSIGHT | FRONTIER 26 Jan 2021

Large-scale threats to society like the impact of climate change and COVID-19 will continue to disrupt society, and lead to, as the International Monetary Fund puts it, "**wartime policy measures**." This will likely trigger extreme actions where **government** agencies use AI and access citizen data in ways that threaten privacy and other civil liberties. Combining "wartime" threats with data-informed capabilities can also lead to data sharing policies and practices biased in favor of quick fixes - for example, the Covid-19 tracking apps inadvertently providing identifying information of infected individuals. Here, the use of data can induce a maladaptive sense of shame, create disgust towards those infected, and ultimately lead to systematic discrimination. In short, recent AI and data-informed policy measures highlight the need for privacy preserving approaches to data-informed policy while also incorporating sufficient domain expertise to minimize biased and unfair interventions.

Preserving Privacy and Debiasing Data-Informed Policy

Fundamentally, machine learning (ML) is about turning data into intelligence that will generalize to new data. There are several steps involved in developing an ML system for a given problem. **Figure 1a** depicts a high-level view of a typical workflow used within industry.

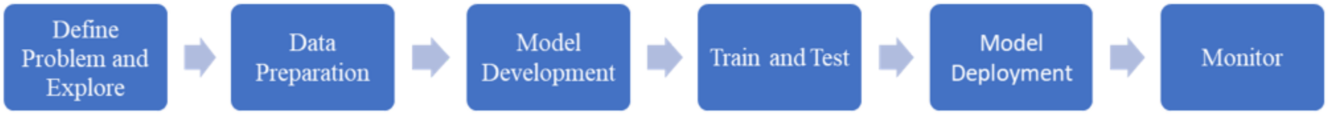

Fig 1a. A typical machine learning workflow.

Though this workflow is standard practice, such an approach is not advisable when it comes to policy because the downstream consequences can be detrimental without explicit steps to preserve privacy and debias policy. Concerns can range from the economic costs of ineffective interventions to significant social costs associated with the erosion of civil liberties. It is therefore important the workflow has added features to account for privacy threats and biased decision-making. **Figure 1b** depicts this modified workflow incorporating four steps to a "machine learning policy support system."

To address the policy *privacy* problem (Step 1), we first introduce the state of the art in privacy preserving machine learning as well as secure and remote computation. We then combine advancements in machine learning and data science with research on human biases to explore the policy *decision-making* problem (Steps 2-4). Along the way we also supply concrete examples of how organizations are using these advancements. Ultimately, we want practitioners and scholars to get a feel for what tools are available and why leaders *need* to use them.

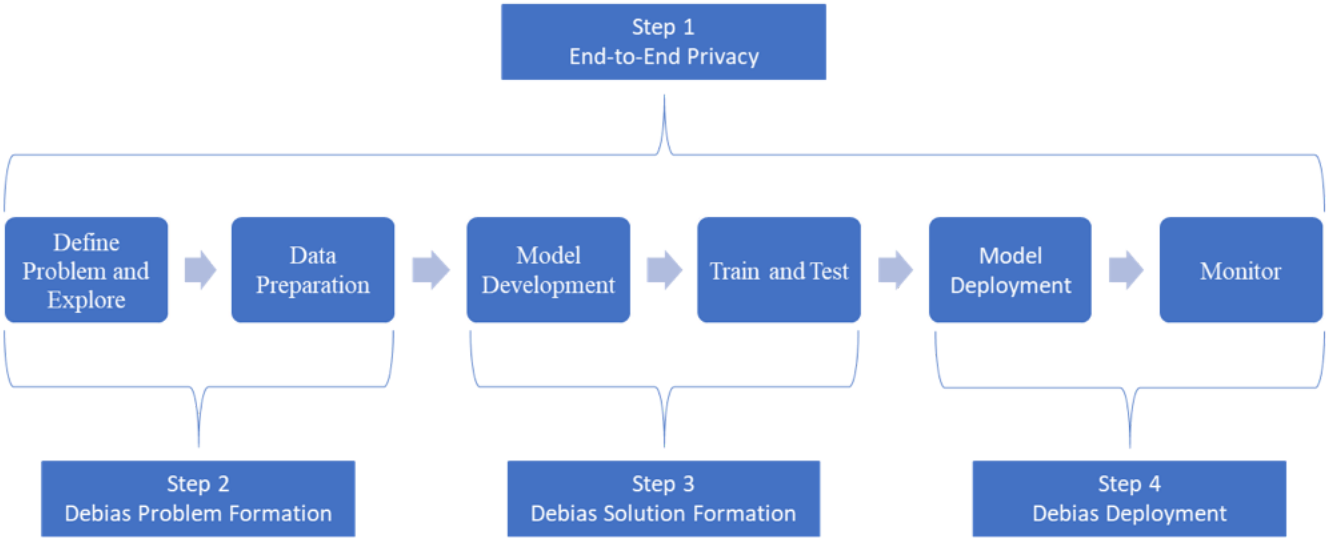

Fig. 1b. A modified workflow for a machine learning policy support system.

Step 1. End-to-End Privacy

Privacy preserving machine learning (PPML) is a policy issue that continues to grow in importance. With relatively new legislation such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), privacy preservation is a topic of interest among policymakers. Such regulation is important because ML models have several stages where malicious actors can access and exploit user data. The areas of risk include (a) within the training data, (b) when inputting user data, (c) model output data, and (d) the model itself. This exposure is obviously a major concern given the pace at which data-informed policy is spreading. In short, growing amounts of personal data require privacy preserving technologies. Below is a sampling of common research and development activities.

Differential Privacy

Differential privacy (DP) is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It enables the collection, analysis, and sharing of *statistical estimates* based on personal data (i.e., patterns in groups) while withholding information about any individual in the data. The intuition here is that DP can "add noise" to user data before it is shared with other parties. When noise is applied, a strict upper bound can be set on the amount of information leakage within any particular query (in some cases 0, in other cases small but acceptable amounts of personal data). Here, leaders can guard against "linkage attacks" where publicly available data overlaps too closely with data held in anonymized datasets (e.g., see the oft-cited example where **a former Governor of Massachusetts had his sensitive medical records re-identified using overlap with his voter registration data**).

Such linkage attacks have in part motivated the U.S. Census Bureau to incorporate DP into their 2020 privacy strategy. Using DP, the Census Bureau can now share and analyze more data than ever before to generate valuable social insights while maintaining privacy. In addition, Microsoft's geolocation system, PrivTree, uses DP to "mathematically blur" the location of any specific individual in their database. Simply put, noise is added to the original (identifiable) data and the subsequent PrivTree output is privatized, making it safe for sharing and analysis. DP subsequently facilitates growth while preserving privacy - thereby reducing the risk of lawsuits, bad PR, and so on. Commercial uses include geographically targeted search engine results, location-based marketing, real-time navigation, and fitness app data sharing.

Secure Computation

Secure multi-party computation (SMPC): Secure multi-party computation (SMPC) is a subfield of cryptography with the goal of creating methods for parties to jointly compute a function over their inputs while keeping those inputs private. The intuition for SMPC is a cryptographic protocol that distributes a computation across multiple parties where no individual party can see the other parties' data. Simply put, multiple parties can securely share sensitive information to gain valuable insights without needing to rely on a third party.

Here is an example: Suppose LinkedIn wants to introduce a feature where a group of users with similar jobs can share their salary data to see where they rank compared to each other. One method LinkedIn could use to carry out this task is to trust a third party to collect the salary data and provide the employees with averages. This, however, is labor intensive at scale and not overly secure (e.g., the third party holding the data is attacked). Rather, LinkedIn could introduce an SMPC protocol where user data is automatically encrypted, the multi-party computation occurs, and the only output the user receives is the average salary. **A recent wage equity study conducted in Boston** used such an approach to assess gender pay gaps while ensuring actual female and male pay totals never left employer servers. Commercially, companies such as Unbound Tech are leveraging SMPC to provide secure computing solutions for many of the world's largest banks and Fortune 500 companies.

Trusted execution environments (TEEs): A trusted execution environment (TEE), or secure enclave, is a secure area of a main processor or chip (i.e., System on Chip) within a mobile phone. Code and data loaded inside are protected with respect to confidentiality and integrity. A TEE has an isolated execution environment providing security features such as "isolated execution." The intuition is that the user data used for ML model training and predictions stays in an environment secured at the hardware level (i.e., it is not accessible by other users).

This protected area can, for example, help leaders **expand democratic engagement by creating a privacy-preserving environment for electronic voting**. It also has significant implications for how individuals manage their medical records and financial information.

Visa, for example, recently **developed LucidiTEE** and **acquired Plaid** (one of the world's largest fintech firms) for $5.3 billion in a bid to capitalize on *secure enclave* capabilities in the payments industry.

Fully homomorphic encryption: **Fully homomorphic encryption (FHE) is a privacy preserving approach where the inputs, output, and intermediate data values are always encrypted**. The intuition here is that executed operations occur directly on the encrypted data instead of having to decrypt first. This allows the developer to implement training or inference without ever seeing user data, and leaders can take advantage of cloud computing (i.e., access to a shared pool of computing resources) while preserving privacy. Any application can (theoretically) be run in this encrypted state, such as efficient energy grids or robust pandemic responses.

Though dataset size is an issue for this computationally intensive process, solutions are emerging for commercially viable applications. The outcomes will add value to many security-relevant tasks such as data sharing, data monetization, and cloud computing - **here are three use cases to consider**. In short, FHE is the gold standard of secure computation.

Remote Computation

Federated learning: **Federated Learning (FL) is a distributed machine learning approach that enables training on a large corpus of datasets in "devices" such as mobile phones while ensuring the data stays local**. FL is a rapidly growing area of research and development. A Google Scholar search for "federated learning" indicates that papers mentioning FL increased five times from 2018 to 2019, and in the first half of 2020 alone, there were 1050 FL papers published.

The intuition here is that the user's data stays in the device and only returns the gradients resulting from local training. The gradients from all the participating devices are securely aggregated into a central server to update a model. The model is then sent back to the edge devices providing improved performance by leveraging updates across the various devices on the network without revealing any of the private information used to train the model.

Leaders can then use data-informed predictions without compromising privacy (e.g., using a federated network of datasets housed in everything from hospital databases to smartwatches for monitoring Covid-19 symptoms and predicting outbreaks). This means that organizations do not need to centralize and amass mountains of sensitive data – thereby exposing themselves to privacy attacks, lawsuits, and so on. It is therefore only a matter of time until many (if not most) sectors adopt this system of collaborative learning.

The Mayo Clinic, for example, recently launched the **Clinical Data Analytics Platform**, allowing a network of participating organizations (e.g., universities, private companies, and government agencies) to collectively train algorithms for improving healthcare outcomes while keeping their stakeholder's data securely on-site. The applications include everything from image recognition for detecting heart disease to pharmaceuticals. The takeaway is that FL allows for unprecedented levels of research and development while upholding ethical standards and privacy.

The Human Factor

Even with this extensive push to solve the *privacy* problem, the above technologies cannot account for many types of human error. Indeed, data and analytics can occur in a secure vault, but if the quality of data and analytics is biased before entering the vault, then the outcome is still flawed. This brings us to the *decision-making* problem.

Data-informed policy incorporating only a narrow and siloed band of social science, for instance, will miss important implications. Covid-19 shaming stemming from hasty policy, for example, contributed to **death threats and possibly even suicide**. Such decision-making policy errors are associated with human cognitive biases. The world is extremely complex, and **humans tend to compartmentalize information to create order and meaning**. It is therefore not surprising policy biases associated with knowledge silos and snap judgements emerge. This is where advancements in ML and data science, now configured for decision support, can help leaders debias policy.

Step 2. Debiased Problem Formation

Natural language processing (NLP) applications such as **topic modeling** and **word embedding** can help leaders better define the problem. A topic model illuminates the topics occurring in a collection of documents while word embedding can transform raw text data into usable insights. A policymaker, for example, traditionally employs analysts to define a problem based on their reading of scientific documents, scouring of social media, and reviewing of news media. Now, NLP advancements can do this at a much larger scale and generate an exceptionally clear picture of the problem.

The intuition is that leaders can use NLP to scan large amounts of information regarding a specific topic (from diverse perspectives) to better understand a problem. For example, why are some experts for and some against the use of tracking and tracing apps? This approach supplies objective insights into how different groups perceive an issue and what topics the problem needs to consider. Such a process helps overcome cognitive biases such as the **confirmation bias** (i.e., the tendency to search for and/or interpret information that confirms one's prior beliefs and/or values) and **anchoring** (i.e., the tendency to focus too heavily on an initial piece of information).

We particularly focus on NLP because expert knowledge across disciplines is often encoded in language. NLP, for example, can capture Covid-19 research from purely qualitative disciplines to written summaries in applied mathematics – thus delivering a richer understanding of the problem. One such tool is COVIDScholar from Lawrence Berkeley National Laboratories. As stated in a recent **news release**, COVIDScholar, "uses natural language processing techniques to not only quickly scan and search tens of thousands of research papers, but also helps to draw insights and connections that may otherwise not be apparent. The hope is that the tool could eventually enable automated science." This debiased, "automated science" allows leaders to explore a broader knowledge space when formulating the problem. In short, it is a guided tour of knowledge.

Step 3. Debiased Solution Formation

Knowledge graphs and **recommender systems** can then help leaders understand where the debiased problem fits within an existing network of knowledge. A knowledge graph is a programmatic approach for modeling knowledge domains and understanding their interconnectedness. A recommender system is also able to filter vast amounts of information to create domains and recommend content. Collectively, they can map the properties of a subject area and suggest choices otherwise overlooked. Thus, instead of biased mental models and narrow solutions, knowledge graphs and recommender systems broaden the solution space.

The intuition is that once a leader accurately formulates a strategic problem using NLP (Step 2), the subsequent key words and phrases defining the problem integrate into a knowledge graph and recommender system to map the problem and generate recommendations for developing a debiased solution (e.g., recommending experts on shame when developing Covid-19 solutions). Further, where typical recommender systems homogenize recommendations, scientists are now combining complex knowledge graphs with recommender systems to broaden the search space, thus leading to more accurate solutions - see this recent article from **Microsoft Research**.

This step helps overcome biases such as **"judgement by prototype."** This is where an individual decides what knowledge is necessary to solve a problem based on salient aspects of its prototype. For example, pandemic problems being prototypically judged as requiring medical professionals and economists while discounting less prototypical aspects of the problem (e.g., knowledge associated with shame and discrimination).

Several companies are using knowledge graphs and recommender systems to counteract such biases. **Thomson Reuters, for example, launched their first knowledge graph in 2017 to help guide policy and practice in the finance sector.** Their knowledge graphs compile a wide range of data about organizations and people (e.g., filings, reports, M&As), allowing companies to better plan their research and more accurately formulate solutions. Similarly, **Ericsson is creating a self-adapting system using a knowledge graph to autonomously develop and implement solutions.**

Step 4. Debiased Deployment

Finally, **ML auditing** works to mitigate any remaining bias and unfairness in deployed models. Leaders, for example, may deploy interventions based on false or skewed assumptions about various demographic groups. So, even if Steps 1-3 preserve privacy, appropriately define a policy problem, and formulate a robust solution, biases hidden in the leader, data, and/or algorithm can still degrade policy.

The intuition is that the audit reduces an intervention's likelihood of undeserving or overserving any specific individual or group. In terms of real-world applications, Aequitas, an open-source ML auditing toolkit, for example, uses a "**fairness tree**" to link different types of fairness with specific real-world policy problems. For instance, the tree specifically asks if the intervention is intended to affect all groups equally or proportional to the group's percentage in the overall population. Such an approach works to mitigate unintended effects (e.g., **Covid-19 policies disproportionately harming ethnic minorities and migrant women**). Here, debiasing incorporates domain experts in areas such as discrimination and inequality. These experts, in consultation with policymakers and data scientists, will help calibrate ethical aspects of data-informed interventions *before* deployment – thus reducing the need for after-the-fact damage control.

Such a process works to overcome instances when leaders overestimate an intervention's efficacy based on specific examples of success without considering an overall failure rate (i.e., ***base-rate neglect***). Incorporating a debiasing audit also helps to avoid policy decisions swayed by **ingroup favoritism** and **ultimate attribution errors** (e.g., promoting a pandemic mitigation policy assuming a group's flawed behavior causes the spread of disease when contextual factors are the underlying cause). Finally, auditing also minimizes a leader's "**bias blind spot**" (i.e., recognizing the impact of biases on others, but not seeing the impact on one's own decision-making).

Step 4 is hugely important because it is the last chance to catch bias before stakeholders start to "feel" policy. The outcome of biased policy touching ground has far-reaching and long-term consequences. Governmental institutions may inadvertently discriminate against certain groups such as the well-documented problems with **ML-informed policy in the criminal justice system**, and firms might rely on biased ML-informed policy in their hiring and promotion practices, **leading to further discrimination** in the workplace. Step 4, accordingly, is *mandatory* quality assurance.

Concluding Remarks

Using technological advancements to preserve privacy and debias policy surrounds leaders with a host of relevant domain experts, engineers, and data scientists. Moving forward, it is thus advantageous, completely feasible, and increasingly important to develop a machine learning policy support system. The four steps we introduce offer a path towards this modern and increasingly beneficial form of governance.

Brian R. Spisak Follow

Brian R. Spisak is a research associate at Harvard's National Preparedness Leadership Initiative, a joint program of the Harvard T.H. Chan School of Public Health and the Center for Public Leadership at Harvard's Kennedy School of Government, and a senior lecturer at the University of Otago in New Zealand. He researches the biological and cultural evolution of leadership as well as the emerging topic of machine learning and leadership. Brian applies his work to issues relevant in business and society, including innovation, sustainability, and the alignment of operational behavior with strategic goals. His work is published in leading outlets including Academy of Management Review, Psychological Science, and The Leadership Quarterly.

Joseph Spisak Follow

Joseph Spisak is the product lead for PyTorch, the open-source AI platform used by Facebook, OpenAI, Microsoft, Tesla, Uber, and many others. His work spans collaborations with many teams, including the AI developer community to bring scalable tools to help push the state of the art forward. Prior to PyTorch, Joseph led a deep-learning product management team and global AI partnerships for Amazon AI. Before that, he was director of the machine learning segment for Intel.

Andrew Trask Follow

Andrew Trask is leader of OpenMined, an open-source community whose goal is to make the world more privacy-preserving by lowering the barrier-to-entry to private AI technologies. He is also a PhD candidate in computer science at the University of Oxford.

References

See in-text hyperlinks.