Data Science Data Governance

Joshua A. Kroll | University of California, Berkeley

This article summarizes best practices by organizations to manage their data, which should encompass the full range of responsibilities borne by the use of data in automated decision making, including data security, privacy, avoidance of undue discrimination, accountability, and transparency.

An increasing number of consequential decisions are made automatically by software that employs machine learning, data analytics, and AI to discover decision rules using data. The shift to data-driven systems exacerbates gaps between traditional governance and oversight processes and the realities of software-driven decision making. And with more and more software-mediated systems turning to machine learning, data analytics, and AI to derive decision rules using data instead of having humans code those rules by hand, this gap in understanding the ramifications of a technical system can exist even for the software engineers, data scientists, and system operators who design, build, deploy, and manage the machines that mediate our modern lives. Whether algorithms are approving credit applications, selecting travelers for security screening, driving a car, granting and denying visas, or determining the risk profile of an accused or convicted criminal, there is a broad societal interest in ensuring the good governance of these technologies and building accountable algorithms.

The dominant position in the legal literature and in policy discussions has moved beyond the idea that transparency will solve these issues but has not set out a full alternative.[1] Disclosure of source code is neither necessary to establish relevant facts for the purpose of oversight nor sufficient to support public or regulatory understanding that enables participation in governance. Further, transparency is nearly always objectionable to those who profit from methods that do not muster protection under patent or copyright. Also, transparency is at times undesirable, as source code disclosure or other detailed knowledge of a system facilitates adversarial activity, such as gaming or exploiting computer systems. In addition, given the important role data play in machine learning, data analytics, and AI systems, source code or other system information, on its own, often does not fully reveal how such systems work. The collection, normalization, exploration, and cleaning of data also affect how systems function.

Businesses that field data-driven systems of all sorts—from the simplest descriptive analytics to the most sophisticated deep-learning models—must further reckon with a thicket of data governance requirements. How was a set of data collected or obtained? Were the data collected directly from customers, or was the dataset purchased from a third party? Is use of the data restricted by a privacy policy or contractual requirements? Do any data protection or other laws

Digital Object Identifier 10.1109/MSEC.2018.2875329
Date of publication: 21 January 2019

apply, and from which jurisdictions? Can these data be combined with other data safely and legally? Are these answers the same for all customers, or do customers in different countries require different policies?

Purely from the perspective of legal compliance, the shift to data-driven decision making presents enormous challenges. Beyond this, there are often ethical and reputational considerations that must be accounted for—no chief executive officer wants to see his or her company pilloried in the media for even perceived discrimination. However, smart data governance policies and practices can confidently navigate these treacherous waters. This article provides examples of the ways data-driven systems can go awry, examines approaches to mitigate and control these problems, and explores how to proactively manage the responsible use of data.

Specifically, new technologies and best practices (both technical and organizational) can support human rights and governance norms to rein in algorithmically driven decision-making systems. As a case study, the EU's new General Data Protection Regulation (GDPR)[2] provides an excellent working example on the details of transparency policies for the governance of automated decision making by data-driven systems. To wit, GDPR Article 22 provides EU citizens with the right to demand that important decisions not be made about them "solely by automated processing." And Articles 13–15 provide rights to notice of data practices as well as a right of access for individuals to data about them. Articles 16 and 17 provide further rights to data correction and erasure, respectively. When applied to data-driven decision-making systems, the GDPR brings into focus many important questions about data governance and calls into question many existing best practices for privacy and data management.

Background

Traditional privacy governance is based around principles of notice and informed consent. But can consent truly be given by a data subject who is not informed about the function of a data-driven decision system? A now-famous story about the retailer Target illustrates the problem: Target discovered a set of 25 products that, when their purchasing was considered together, would reliably predict whether a customer was pregnant.[3] Marketing could be directed to such customers aggressively in a pregnancy-specific way, given that data show that consumers' purchasing habits are likely to change with the arrival of a baby, meaning that pregnancy is a prime opportunity to acquire new regular customers. And so, an enraged father came to brandish a circular for baby products at the manager of his local Target store, mistakenly believing that the company was encouraging his teenage daughter to become pregnant without realizing that she already had. Clearly, neither the mother nor her father had consented to this prediction, making it ethically problematic even though it was correct.

Although one might argue that traditional purpose restrictions on the use of data, applied at the time of collection, would prevent such problems, the sensitivity of the prediction is difficult to predict from generic disclosures. Whereas many consumers would likely object to a notice that says, "Data about your purchasing habits may be used to predict medical conditions including pregnancy," few in practice object to the more pallid, "Data about your purchasing habits may be used for marketing purposes and to suggest products you may be interested in."

Further privacy risk in this scenario (both to consumers, who wish to control their personal information, and to data controllers, who may inadvertently gain more sensitive data than they planned to) comes from the fact that personalized marketing can, in some cases, reveal the purchasing habits of consumers (which, in turn, can reveal sensitive or protected medical conditions, such as pregnancy). In a study of online retailers and recommendation systems, researchers were able to demonstrate how to use changes in product recommendations to identify the purchases and ratings of other customers and users.[4] The traditional data governance ideas of effective notice, informed consent, and restrictions on the purposes for which data may be used do not readily apply in this scenario.

Data Governance Best Practices

The fastest way forward on questions of responsibility in data science is through the development of best practices for data governance in the age of machine learning and big data. Although several statements of data governance principles exist and have been adopted by professional organizations or standards bodies (see later), it is unreasonable to imagine that every organization that wishes to use data-driven approaches will engage in a full first-principles analysis of their behavior. Smaller organizations without dedicated resources for such detailed review benefit especially from the ability to rely on the shared judgment of an industry writ large. Still, data responsibility questions are highly contextual, and solutions do not carry well from one situation to another. This article translates emerging practices as well as suggestions from statements of principles and legal requirements into concrete recommendations that can be taken up by organizations using data science and machine learning. Emerging best technical practice and new regulatory guidance will shape the future of even existing governance mechanisms to address these problems.

Treat Information Security and Privacy as First-Order Problems

First and foremost, a responsible data governance strategy must incorporate strong strategies and programs in both information security and privacy. Any data collected and retained pose some risk of breach, and by far the simplest approach is to limit the data collected to only what is absolutely essential and avoid retaining data once they are no longer business critical. Whatever data remain that must be retained (for example, lists of active customers and their billing information) should be secured from outside hackers as well as deliberate misuse by insiders. Retained data should be associated with metadata that identify provenance, sensitivity, and known legal or contractual limits on uses. (For example, commercially procured data may only be available for certain uses or may be restricted in terms of what other data they can be combined with for purposes of analysis by contract with the vendor. Laws such as the Fair Credit Reporting Act in the United States prohibit the use of certain factors in credit decisions or decisions based on background checks, such as employment decisions, even when collecting such sensitive information is legal.)

Data should be encrypted at rest and in transit, with encryption keys subject to suitable access control mechanisms so that sensitive customer data are available only when and where they are needed and by employees with a verified need for access. Sensitive data stores should be monitored so that queries against them can be audited later and correlated to an approved business need. As the litany of major recent data breaches and privacy failures shows, a firm information security posture and a robust privacy program are the foundation for any responsible data governance strategy.

Minimize Data Collected and Retained; Scrub or Aggregate Retained Data When Possible

Similarly, when data must be retained, data practitioners should consider whether they can be scrubbed to a lower level of sensitivity. For example, whereas keeping raw visitor logs is convenient for a website-hosting company, customers are likely only interested in the count of visitors or perhaps the count of visitors stratified by some attribute, such as geographic region, type of client device and software, or client network operator. And system operators are unlikely to need technical debugging information after a day or two (longer-term requirements for operational data retention could be approved on an exceptional basis or covered under programs for aggregating specific kinds of data, such as all requests from a particular network or type of device).

Moreover, raw logs present a major risk to privacy not just from hacking but also from legitimate requests. Law enforcement, for example, often requests the logs of service operators to identify persons of interest in ways that implicate the privacy of users unrelated to the investigation. But data that are not present or that have been aggregated or scrubbed so as to no longer confer the information of interest cannot be produced. Understanding how and why data must be retained and how they will be used can inform when data can profitably be discarded, properly anonymized or pseudonymized, or otherwise transformed to minimize risk and sensitivity.

Consider the Risk of Identifiability

When analyzing data for sensitivity to ensure that minimal data are collected and retained and that retained data are aggregated or scrubbed when possible, it is tempting (and often necessary under legal regimes or standard compliance certifications) to identify certain categories of data as personally identifiable information that must be treated safely and to treat other data as nonsensitive and thus safe to store and process. However, research has shown that this distinction is a fallacy.[5] Nonsensitive data are often unique enough to a person that they can prove identifying under the right circumstances. Responsible data governance must consider the risk that retained data could be reidentified (possibly using data that come from elsewhere) and should consider whether formal guarantees like those from differential privacy are useful to safely retaining and analyzing data. Re-identification risks will depend on the type of data in question and the context of the process or system in which it is being used.

Designate and Empower a Data Use Review Board

Data scientists must constantly question the responsibility of their methods and findings and be prepared to forgo analyses that would violate laws, privacy norms, contractual requirements, or customer trust. To assist in this, companies employing data science should designate review boards and empower them to approve or deny the collection of new data (for example, "Should we retain purchasing habits of our customers, and for how long?"), the investigation of sensitive questions using company data ("Can we predict pregnancy from purchasing behavior?"), and the deployment of insights from such analysis ("Knowing that we can predict pregnancy from purchasing behavior, should we produce personalized marketing materials based on this prediction?").

Such review boards should contain stakeholders from many functions, including data science, information security, legal, compliance, marketing, and any other key functions that support understanding customer

relationships. The more diverse such a board is in terms of the backgrounds and functions of the people who comprise it, the more likely it will be to uncover problems quickly and the more valuable its insights will be to the data science process and the enterprise as a whole. Boards should focus on responsible data use by examining the details of data collection, normalization/cleaning, analysis, and use, attempting to foresee how data use will affect customer trust, the company's reputation, the company's risk of data breach, and the company's risk of legal noncompliance as well as how approving or denying the activity will affect the company's core product and service offerings. Such boards should not hesitate to seek the advice of outside experts, who may be better equipped to foresee relevant problems, or to convene a (suitably representative) panel of trusted customers to examine reactions. Some data activities may require study and a formal impact statement before they can be responsibly approved.

Write and Publish Data-Focused Social Impact Statements

Impact statements provide a formal, structured process to investigate whether foreseeable issues exist in data collection and analysis practices and provide a digestible view of the risk of data processing in specific cases. Entities concerned about the equities of data analysis should include such concerns and any mitigations adopted in their privacy impact statements and should consider producing similar social impact statements for data-driven systems and processes. Impact statements provide critical transparency about the organizational acknowledgment of risk and the techniques used to mitigate those risks, without foreclosing any specific activities up front.

Environmental impact statements demonstrate the usefulness of such work in engaging stakeholders and providing a record of decisions and the factors that supported them. Further, impact statements provide an analytical means to consider the tradeoffs of certain approaches (perhaps the same insight could be gained through a more palatable and responsible use of data in a particular case—for example, the threshold for making a particular prediction will likely trade off false positives and false negatives, and the difference may be relevant to customer trust, as with Target's pregnancy prediction score). Such statements could be purely for internal consumption, focused on convincing organizational leadership that data analysis is valuable and does not create undue risks while providing a record of issues considered for review by auditors. Statements could also be published to help engage the trust of customers and civil society groups or as a way of soliciting feedback about the impact of data-driven decisions.

Attempt to Explain Data-Driven Processes to Breed Understanding

To trust data-driven decision processes, data scientists and decision subjects alike must understand them. Understanding is supported by the ability to explain what the process is doing and how it reached its decisions. In many cases, it is important to understand both how *specific* decisions were made (that is, to account for the factors in any particular decision, which is sometimes called a "local explanation") and the rules in force *in general* (that is, to understand which factors are at play in all decisions of a certain type or from a certain system, sometimes called a "global explanation"). Both kinds of understanding are served by data analysis methods that privilege explainability. Many such methods exist today, and the question of what constitutes a useful explanation is an active and exciting research area. (For surveys in this rich area, see Guidotti et al.[6] and Doshi-Velez and Kim.[16]) However, explainability is neither a cure-all nor an unalloyed positive.

Although explanations can help system developers, decision subjects, and nonusers alike to understand when data-driven processes are trustworthy and correct or to challenge incorrect assumptions or faulty methods, it is important to remember that explanations alone do not create understanding. Explanations must be supported by sufficient other evidence to be believable, should explain the causes of an outcome when possible, must be targeted to the people meant to receive them, and must adequately engage the task at hand. Otherwise, explanations risk giving credence to an incorrect model. A good explanation for an incorrect decision might even lead people encountering a model to discount its likelihood for making additional incorrect decisions. Comparatively little research has been done on the human factors of explanations or the situations in which they are appropriate.

In addition to providing explanations, data-driven systems can be made more transparent through the disclosure of analysis methods and the underlying datasets, when possible. The release of data must be carefully considered, though—in addition to concerns about proprietary advantage and privacy, datasets are often very sensitive to the particulars of their collection context and methodology, and released data risk being repurposed without consideration of these factors. Thus, released data should always be accompanied by information about provenance and processing. Data consumed from outside an organization rather than carefully collected should be evaluated for fidelity to the phenomenon under consideration.

Consider How to Support Ongoing Auditing of Correctness and Challenging of Assumptions

It is not enough to consider the correctness of a data-driven system a priori or to evaluate correctness as the system is being designed or fielded. Interrogation into the fidelity of a data-driven system is an ongoing effort, impelled by the twin risks of *modeling error*, the taking on of unwarranted assumptions by way of choosing how to describe data and missing details of the world, and *concept drift*, changes in the world that can invalidate assumptions baked into collected data or the data collection and normalization methodology. It is important for anyone making decisions based on data to understand the assumptions baked into the data and to ask if those assumptions are warranted by reality (or at least an organization's best understanding of it).

Data scientists must find ways to validate their predictions continually and should plan to monitor the performance of their systems well after launch. A useful method originating in the social sciences is *auditing*, which asks how a system would behave on differential inputs (for example, does a decision process rank similar resumes at similar levels when the resumes appear to be from applicants of different genders or races?). Audits are a form of black-box testing designed to validate and support the conclusions of a system or determine in what ways those decisions might be incorrect or unfair. The results of an audit may be intended for the data scientist alone, superiors in his or her organization, or the public. Data governance strategy should also consider when it is appropriate or even necessary to facilitate external audits by trusted academics, journalists, civil society groups, or even by the public at large. When public auditing is useful, systems could be modified to support querying on synthetic data and demonstrate how output would have changed under the hypothetical situation were the input slightly different.

Other, stronger forms of testing and validation should be considered as well, including white-box testing methods in which the structure of the model figures in the testing. For example, data scientists may find it useful to treat different classes of inputs under different regimes, and testing should take account of the validity of each component in addition to the whole. Concretely, suppose that Target used different models to predict the future purchases of regularly returning customers and occasional customers: it is important to understand the correctness of these segments individually, in addition to understanding how well aggregated future purchase behavior is being predicted.

> **"There is a broad societal interest in ensuring the good governance of these technologies and building accountable algorithms."**

Look for Systematic Biases in Data, and Consider Potential Causes of Unfairness

Systematic bias can enter datasets and data analysis methods at all levels: data may be collected from a nonrepresentative sample, such as when data come from a source of convenience (say, photo datasets culled from social network posts, which represent only the subset of users who post photos and could leave out or skew against privacy-conscious minorities or users of lower socioeconomic status) rather than a source that reflects the universe of possible values of interest. Data may also be subsampled or coalesced in a way that disadvantages particular groups more than others (such as if data about employee performance are grouped and thresholded by employee tenure, and women have a shorter average tenure, dropping them out of the longest-serving cohorts). Further, data may be skewed by human interaction, such as labeling outcomes, handling missing values, pruning outliers, defining groupings, or encoding categorical variables—all context-specific problems for which solutions can make or break a model.

Human data scientists define the problem to be solved by data analysis, choosing things like which particular methods work best, how to measure success, what values to optimize for, and how to select parameters and hyperparameters. Systematic bias in the data or the data analysis method can easily cause systems to treat different subgroups differently and in some cases could rise to the level of formal illegal discrimination based on a legally protected attribute. When the value being predicted is not easily measured directly, such as the risk of future behavior (for example, predicting whether a set of financial transactions is indicative of terrorist behavior), these problems become even trickier, as there is not a good way to identify when disparities in predictions constitute meaningful unfairness. It is particularly important to

rule out such behavior from black-box models, such as random forests and deep neural networks, which may infer protected attributes as new features on which to base classifications.

The particular way data are considered matters as well—patterns that exist in aggregated groups may disappear or even reverse when those groups are considered as separate subgroups, due to a phenomenon called *Simpson's paradox*. For example, graduate admissions data from the University of California, Berkeley, in 1975 appeared to show a significant bias against women when considered across the university, but analysis by department showed no such disparity and even a small bias in favor of women.[7] This is because women applied in greater numbers to more selective departments than men, causing a greater fraction to be rejected overall.

As with the problem of identifiability, the naive approach of simply removing the sensitive attributes from the data will not solve the problem. Sensitive attributes are often encoded in a latent manner in other nonsensitive data (for example, ZIP codes in the United States correlate closely with race, cultural heritage, and socioeconomic status). This implies that mitigating for systematic data biases must be an affirmative step deriving from an understanding of the nature and source of the bias as well as the best way to respond to it. Fortunately, many techniques exist to provide invariant guarantees that data analysis does not pick up certain types of bias, although how to build systems that apply these is an area of active research.[8,17,18]

Thus, it is important to remain vigilant and test systems for bias both during development and after they are fielded. Auditing, especially by groups that may be affected by bias, is also critically important to investigations of unfairness. Unfairness can be difficult to define during development, so the ability to engage with a system interactively and challenge or change a decision later can also avoid unfairness in many situations. Finally, evaluations of unfairness must consider how the spoils of data analysis will translate into real-world actions.

Consider Not Just Successes but Errors and Feedback Loops as Well

Responsible data governance considers not only the fairness of correct predictions but of errors as well. What makes a decision fair or equitable is not just whether it considered an appropriate and relevant set of factors or didn't consider proscribed or protected factors, or even whether the decision process is always correct. It matters that mistaken decisions do not disproportionately harm individuals or protected groups. Consider the story of COMPAS, a model that uses criminal history and a behavioral interview to predict the likelihood that an individual will recidivate (proxied through rearrest) during a two-year timespan. ProPublica argued that the way COMPAS was deployed in Broward County, Florida, was biased against African Americans, finding that it rated them as a high risk (when they were not rearrested) more than twice as often as it did for white people.[9]

False positives are important in this setting because they are likely to cause the putatively high-risk individual to be treated more harshly by the criminal justice system, receiving higher bail or pretrial detention or even a harsher sentence postconviction. Each of these contributes to assigning individuals a higher recidivism risk, implying that errors in the model cause a negative feedback loop in which individuals misclassified as high risk may in fact become high risk as a result of the misclassification! (Many similar feedback loops are described in O'Neil.[10]). Northpointe, the firm that created and marketed COMPAS, countered that COMPAS has equal predictive parity for both African American and white arrestees (meaning that, given an individual's COMPAS score, the chance that he or she will be rearrested within the following two years is independent of race). However, this property, coupled with the fact that African Americans are arrested at a much higher rate in Broward County (and many jurisdictions in the United States), implies mathematically that the distribution of false positive errors must fall unequally on African Americans.[11]

Design Systems to Allow the Challenge and Correction of Incorrect Decisions by Humans

Although the goal of automated data-driven decision making is to support speed and scale, allowing intelligent decisions faster or more cheaply than decisions by humans given case by case, it is critical that responsible data governance define mechanisms for the outcomes of analysis or automated decisions to be challenged. Not only should there be externally visible mechanisms to engage with automated results, an internal role should be designated that is responsible for owning the outcomes of the automated process and of the human-mediated escalation process. This role should be prepared to deal with both individual-level and broader societal-level claims of unfairness about the automated system. And some role must be designated as directly accountable for problems in the operation of data-driven systems.

Further, systems should produce enough evidence as they are operating to allow decision subjects to determine whether decisions were correct and allow a review process to determine precisely what

happened and why. Indeed, whereas decisions taken by people are subject to the whims of humans, decisions made in software by machines can be made fully reproducible. Careful review of the actions of a properly designed data-driven decision-making system can always determine whether those actions were correct and intended.[20] Systems must be designed to facilitate this kind of oversight process.

A difficulty in planning for review by humans is understanding how to manage it while capturing the speed and scale benefits of data analysis. However, this is a tradeoff and, like any customer support problem, can be managed through careful consideration of what triggers escalations from software to humans. Further, the cost of escalations to a human review incentivizes the development of high-fidelity decision-making processes, closing the feedback loop of investigating model correctness and accuracy. It is important to track the situations where review was necessary and feed that back in to the development of the decision process to improve the automated portions of a process. In addition, such data facilitate examining whether the override process is being abused to favor or disfavor particular individuals or groups.

Data Responsibility Principles and Emerging Standards

In the service of advancing responsible data governance and fairness, accountability, transparency, and responsibility in data-driven systems, many organizations have proffered either statements of responsible data principles or standards against which to measure systems. Because these normative concepts are highly contextual and depend on the nature of the system at issue, the kinds of data or decisions being considered, and even the nature of the populations to which these systems are applied, these documents are best thought of as guideposts or frameworks for evaluation and measurement rather than formal, technical standards. However, such structured evaluation is useful for developing a responsible data governance program, and we summarize several such documents here to synthesize their similarities and differences.

Dagstuhl Principles for Accountable Algorithms

In 2016, a group of academics from a variety of fields came together and produced a framework with light-touch guidance for building what they call "accountable algorithms" as well as a draft social impact statement for data-driven technologies. They note that "accountability in this context includes an obligation to report, explain, or justify algorithmic decision making as well as mitigate any negative social impacts or potential harms."[12] The principles rely on five major pillars—responsibility, explainability, accuracy, auditability, and fairness—each of which comes with a short statement of advice describing its meaning and a set of guiding questions designed to enable practicing data scientists to implement each pillar meaningfully. The pillars and guiding questions are "purposefully underspecified" to allow them to be "broadly applicable."[12] The authors propose that social impact statements for algorithms, at a minimum, address all five pillars and the associated guiding questions.

ACM US Policy Council Principles for Algorithmic Transparency and Accountability and Professional Codes of Ethics

The US ACM Policy Council produced a short statement defining algorithms, analytics, automated decision making, and risks associated with these processes. They present a set of principles to guide responsible "algorithmic transparency and accountability" meant to be supported by the ACM Code of Ethics. These seven principles include awareness, access and redress, accountability, explanation, data provenance, auditability, and validation and testing.[13] These map approximately onto the best practices described in the Dagstuhl principles. Each principle is accompanied by a short description aimed at how policymakers could implement it as a requirement. However, these descriptions also provide a useful inspiration to the development of a responsible data governance program.

Center for Democracy and Technology Digital Decision Project

As a tool for system implementers and policymakers considering requirements for responsible data-driven automated decision-making systems, the Center for Democracy and Technology started the Digital Decisions project.[14] The project's major report provides case studies from a wide swath of industries to identify the "prevailing policy principle" around each of a number of issues. Specifically, the report identifies four major principles around which responsible uses of data can be organized: fairness, explainability, auditability, and reliability. The project also provides a tool meant to interactively guide system designers through a process of considering key questions to evaluate their consistency with the four major principles articulated in the report.

The last of the four principles, reliability, is the most distinct from statements of principle found in other documents and complements the well-researched triad of fairness, accountability (supported here by

auditability), and transparency (supported here by explainability). The report defines reliability as the property that "a system must be able to be trusted to behave as is expected and anticipated by its designers. The ways in which it deviates from these expectations can be monitored and addressed."[14] As presented, reliability is simply the baseline requirement that a system be well specified and correct with regard to that specification. This is less a best practice and more a baseline requirement of a functioning system, but it underscores the best practice that systems be interrogated on an ongoing basis for correctness and validated against their designers' best understanding of the real world in which they operate. Further, it demands careful thinking about how correctness is defined, realized, and measured.

IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

The IEEE Global Initiative aims to bring together top professionals working on AI and other data-driven autonomous systems to build consensus around timely issues and support ethically aligned design that adequately supports human factors and human values.[15,19] The initiative believes that capturing moral values and principles is what is necessary to make autonomous technology deployable in the real world and fully benefit from its potential.

The initiative's flagship report, "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems," defines a grand vision for the design of autonomous systems that support human values. The report is arranged around three major principles: "1) Embody the highest ideals of human rights that honor their inherent dignity and worth; 2) Prioritize the maximum benefit to humanity and the natural environment; and 3) Mitigate risks and negative impacts as AI/AS [autonomous systems] evolve as sociotechnical systems."[15,19] The report aims to identify issues that could violate these principles and provide candidate recommendations to policymakers and system designers. Although the document focuses largely on issues with safely constructing AIs, it does suggest principles for the management of "personal data and individual access control" focusing on data ethics and the importance of personal data control.

IEEE Standard P7003—Algorithmic Bias Considerations

The IEEE, via its Algorithmic Bias Working Group and under the auspices of the Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, is also running an ongoing standards-setting process to define certification for "accountability and clarity around how algorithms are targeting, assessing, and influencing the users and stakeholders of said algorithm."[15,19] The goal of this standard is to provide a certification that "will allow algorithm creators to communicate to users, and regulatory authorities, that up-to-date best practices were used in the design, testing, and evaluation of the algorithm to avoid unjustified differential impact on users."

Although the standard provides for methods to counteract "negative bias" (that is, biases that are unwanted, as opposed to the innocuous meaning of "bias" in statistics, namely, the distinguishing that data-driven models are intended to make among patterns in the data), the standard does not define an explicit compliance regime or define best practices for building data-driven systems in a positive way. Rather, the standard defines a number of benchmarking procedures for establishing data quality and guidelines for determining how models, once built, should be applied to minimize concept drift. Finally, the standard approaches the question of how to manage the interpretation of model output, which is a very underexplored but important area for investigation. The IEEE is running a number of other standards processes surrounding AI ethics and safety, especially around the behavior and safety of autonomous systems.

Data Ethics, Responsible Data Governance, and the GDPR

Armed with the consensus view of how to build responsible data governance, we can turn to examining the most up-to-date policy regime for handling data-driven systems, the new GDPR in the EU, and consider how well it captures or encourages the best practices and principles discussed previously. We do not intend to describe what the GDPR requires or prescribe any specific compliance regime, only to evaluate the extent to which the rule might justify ideal behavior and improve responsible data governance.

The GDPR is a unified rule, effective in all EU member states beginning 25 May 2018, and it repeals and replaces the prior Data Protection Directive (Directive 95/46/EC), which instructed each member state to put into force a national data protection law meeting minimum requirements. The GDPR is widely understood to apply to the data processing of all EU citizens, whether that processing is performed by an EU entity or a foreign entity. Following Brexit, the United Kingdom adopted the GDPR as its local data protection law to enable cross-border dataflows with its many EU trading partners. The law is organized into 99 binding Articles, which

define the operative text. Accompanying the Articles are 173 Recitals that, although they are nonbinding, provide background and interpretive guidance for the law.

Taken together, Articles 13 and 14 provide a strong right of notice for personal data processing, requiring that data subjects be notified about any data processing that concerns them, including a designation of the responsible data controller, the period for which the data will be retained, the existence of automated decision making, "meaningful information about the logic involved" in automated decision making, and a notice of whether data will be used for "a purpose other than that for which the personal data were collected" as well as a specification of further processing purposes. These rights apply whether data are collected directly from the data subject (Article 13) or originate from another source (Article 14). For data-driven systems, the right of notice corresponds to questions about transparency and explanation—for example, under what circumstances does a model constitute personal information about an individual data subject? Time and regulatory practice will tell, although the meaning of "meaningful information about the logic involved" in data processing is the subject of intense scrutiny by academics, policymakers, and lawyers alike. Certainly, our suggestion that data-driven systems support auditing is relevant here, as are questions of understanding the nature of biases in data and the distribution of errors. Data governance processes should be prepared to answer detailed questions about each of these.

> **Data scientists must constantly question the responsibility of their methods and findings.**

The question of what constitutes a sufficient designation of the purpose or context of data processing is also difficult, especially given the unexpected and far-reaching predictions often encountered in data analysis (recall the example of the Target pregnancy prediction score). Here, questions of purpose are well supported by analysis of foreseeability undertaken by ethical review boards and as part of the impact statement process. Indeed, the GDPR requires "data protection impact statements" in certain cases, and such statements will be more effective when they engage with questions of data governance and bias in addition to privacy. Further, such careful analysis during the design of a system can help more clearly define the outline of what it means for a system to be operating correctly and provide organizational visibility for the question of what outcomes are and are not intended.

Article 15 provides a right to data access by data subjects, meaning that individuals may request data about them from data processors. This implies the practice of having a responsible oversight party and a plan for explaining the actions of data-driven systems. However, the extent to which a model, once trained, constitutes personal data about a subject will be an interesting area of practical rulemaking, especially as the fields of model inversion and model reverse engineering improve the techniques available to adversaries who wish to extract private personal data from models or decision outcomes.

Articles 16 and 17 provide rights to the rectification and supplementation of personal data and the erasure of personal data, respectively. Both of these "editing" rights demand clarity from the data governance process about where data about a subject are located, how they are fed into models, and what roles within an organization are responsible for conflicts concerning data accuracy. In addition, the threat of subjects demanding their data be corrected should provide an incentive to businesses to ensure the correctness of the data up front, as it is likely significantly cheaper to work with data that are correct from the outset than it is to process many one-off claims from individuals.

Finally, Articles 21 and 22 provide data subjects with a right to object to the processing of their personal data and a right to demand that decisions be made about them "not solely on the basis of automated processing." Although these naturally support our proposed best practices of allowing human-driven redress, review, and override of decisions, they also present a compelling argument for ensuring that models are trustworthy to data subjects (to limit objections to a reasonable number). Trustworthiness is supported by all of the practices suggested previously, but in particular those related to security, auditing, the assessment of impact, and review by a competent review board. In particular, the right to demand that a subject's data "no longer be processed" presents interesting questions for how to handle models trained using that subject's data. (Recital 69 suggests that the burden of trading off "legitimate interests" of the controller in continuing to use the subject's data lies squarely with the data controller, although no formal opinion on this question yet exists, and it will be another area of practical rulemaking open to interpretation as the law goes into effect.) Responsible data governance practices must be prepared to receive and handle objections to automated processing and, when necessary, substitute human decision-making processes while ensuring process fairness and equitability.

Today's data scientists have the opportunity to lead, building responsible data science practices and

machine-learning systems that can be trusted by ordinary consumers and that will define best practice long into the future. In doing so, practice can shape policy implementation in the EU and beyond, steering public discussion away from fears of bias, lack of accountability, and loss of control. The full promise of data-driven systems can be realized only when regulators and the public believe these systems are built and operated responsibly. The practices suggested in this article are a firm next step in that evolution. Organizations, whether public or private, must consider how they will build and field responsible data governance regimes in the coming years, especially as so many of the hallmarks of responsible data governance are becoming legal requirements around the globe. ■

References

1. J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," *Univ. Pa. Law Rev*, vol. 165, no. 3, pp. 633–706, 2016.
2. European Parliament. EU 2016/679. (2016, Apr. 27). Regulation of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en
3. C. Duhigg. (2012, Feb. 16). How companies learn your secrets. *NY Times Magazine*. [Online]. Available: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
4. J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, "'You might also like': Privacy risks of collaborative filtering," in *Proc. IEEE Symp. Security and Privacy (SP 11)*, 2011, pp. 231–246.
5. A. Narayanan, J. Huey, and E. W. Felten, "A precautionary approach to big data privacy," in *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection*, S. Gutwirth, P. De Hert, and R. Leenes, Eds. New York: Springer, 2016, pp. 357–385.
6. R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. (2018). A survey of methods for explaining black box models. arXiv. [Online]. Available: https://arxiv.org/abs/1802.01933
7. P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975.
8. S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 2125–2126.
9. J. Angwin, J. Larson, S. Mattu, and L. Kirchner. (2016, May). Machine bias: Risk assessments in criminal sentencing. ProPublica. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
10. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Portland, OR: Broadway Books, 2017.
11. J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. Innovations in Theoretical Computer Science Conf.*, 2016, pp. 43:1–43:23.
12. S. Abiteboul, G. Miklau, J. Stoyanovich, and G. Weikum. (2016). Data, responsibly (Dagstuhl seminar 16291). [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2016/6764/
13. ACM U.S. Public Policy Council. (2017, Jan.). Statement on algorithmic transparency and accountability. [Online]. Available: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
14. A. Lange. (2016). Digital decisions. Center for Democracy and Technology. [Online]. Available: https://cdt.org/issue/privacy-data/digital-decisions
15. R. Chatila and K. Firth-Butterfield. (2017). The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. [Online]. Available: https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
16. F. Doshi-Velez and B. Kim. (2017). Towards a rigorous science of interpretable machine learning. arXiv. [Online]. Available: https://arxiv.org/abs/1702.08608
17. Conference on Fairness, Accountability, and Transparency. (2018). [Online]. Available: https://fatconference.org
18. Workshop on Fairness, Accountability, and Transparency in Machine Learning. (2018) [Online]. Available: https://fatml.org
19. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. Version 2. [Online]. Available: https://ethicsinaction.ieee.org/
20. J. A. Kroll, "The fallacy of inscrutability," *Phil. Trans. R. Soc.*, no. 2133, 2018. doi: 10.1098/rsta.2018.0084.

Joshua A. Kroll is a computer scientist and postdoctoral research scholar studying the governance of technology at the University of California, Berkeley, School of Information. His research interests include studying how technology fits within and shapes its human-driven, normative context and upholds such values as fairness, accountability, transparency, and ethics. Kroll received a Ph.D. in computer science from Princeton University, where he was awarded the National Science Foundation Graduate Research Fellowship in 2011. Contact him at jkroll@jkroll.com.