Review of Futures Markets (RFM)
ISSN: 1993-7116
Volume 32 Issue 1 (2024)
Website: https://fra.st

Government Oversight of Data and Responsible AI: Bringing Together Business Strategy, Technology, and the Law

Prof. Octavian L. Mercer
Centre for Digital Governance, Stonemere University, Edinburgh, United Kingdom
Dr. Elisa F. Navarro
Faculty of Law and Technology, Instituto Superior de Innovación Jurídica, Valencia, Spain
Dr. Joon-Ha Min
School of Information Policy, Hanseong Advanced University, Busan, South Korea

Submission: 25.02.2024; Acceptance: 04.04.2024; Publication: 24.06.2024

Abstract
Technologists, law scholars, and business strategists all care about data governance and responsible AI practice because artificial intelligence (AI) is quickly becoming a part of everyday business tasks. This article combines ideas from research on information systems, comparative regulatory analysis, and management studies to look at how businesses can use responsible AI in a world where rules are becoming more scattered. We look at how data governance has changed over time from keeping records for compliance reasons to a strategic, lifecycle-based field. We also look at the main ethical frameworks that have shaped the conversation about responsible AI and the legal architecture that is emerging around tools like the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act. We then look at how these technical and legal needs affect organizational strategy by using examples from the public sector, healthcare, and the financial sector. At the end of the piece, it suggests an integrated framework that sees data governance, algorithmic accountability, and business value creation as goals that support each other instead of being in competition with each other. It also lists research and policy questions that need to be answered in the next ten years.

Keywords: data governance, responsible AI, algorithmic accountability, AI regulation, GDPR, EU AI Act, digital ethics, business strategy

1. Introduction

AI is no longer just something that is studied in a lab. It is now used in banking, healthcare, shopping, transportation, and public administration. A change that is just as important but not as obvious has made this diffusion possible: the gathering, combining, and selling of huge amounts of personal and business data. Many of the predictive and creative abilities of modern AI systems come from data. The quality, provenance, and control of that data now play a big role in deciding how trustworthy, fair, and legally defensible an AI system is. So, data governance and responsible AI are not two separate fields that sometimes cross paths; they are functionally linked. Bad data

governance leads to AI that is not reliable, fair, or follows the rules. On the other hand, bad AI governance can make a well-managed data farm a source of legal and reputational risk.

In real life, this interdependence has been hard for groups to handle. IT and records management experts have traditionally been in charge of data governance, which focuses on data quality, security, and keeping regulatory records. Responsible AI, on the other hand, comes from a different set of ideas that include applied ethics, computer science fairness study, and human rights law. Business planning, on the other hand, has mostly seen both areas as places to cut costs or lower risks, not as places to gain a competitive edge. There is a persistent gap between the aspirational principles spelled out in corporate AI ethics statements and the operational practices that shape how systems are built, deployed, and monitored, as has been seen in many different organizational settings (Mittelstadt, 2019).

This article fills in that gap by combining three different types of literature: the literature on information systems on data governance maturity; the literature on data protection and AI regulation; and the literature on management on how to strategically implement ethical commitments. Our main point is that responsible AI can not be reached just by making technical fixes, like bias-mitigation algorithms or model cards, or by following the law all the way up, like doing data protection effect assessments. If you want to be responsible, you need a governance architecture that covers the whole lifecycle of the data and models, is built into the way organizations are set up and the incentives they offer, and sees regulatory compliance as a floor rather than a ceiling.

The rest of the article is organized like this. In Section 2, we look at how data governance has changed over time from a compliance function to a strategic, lifecycle-based field. In Section 3, we look at the main ethical frameworks that support responsible AI. We also look at their well-known flaws when used without accountability systems. In Section 4, we look at the new laws that are being made around the GDPR and the EU AI Act, as well as the problems that come up with moving data across borders. In Section 5, we look at how these requirements are reflected in the strategy, governance, and performance metrics of the company. In Section 6, examples from the financial services, healthcare, and public sectors are used to show how these changes happen. In Section 7, we talk about ongoing problems and tensions, and in Section 8, we suggest a unified approach that should help with both research and practice.

2. The Evolving Landscape of Data Governance

2.1 From Compliance to Strategic Asset

Data governance usually means having authority and control over how data assets are managed. It includes the rules, guidelines, roles, and procedures that make sure data is correct, accessible, safe, and used correctly (DAMA International, 2017). Data governance used to be the same thing as recordkeeping and regulatory compliance. It meant making sure that financial data met audit requirements, customer records were kept for as long as the law required, and people could not get to sensitive information without being logged in and given permission. Sector-specific rules in

finance, healthcare, and telecommunications reinforced this focus on compliance by seeing data mostly as a liability that needed to be managed rather than an asset that could be used.

This way of thinking has been turned upside down by the rise of large-scale data and then machine learning. A lot of people now see data as a strategic tool that can be used to make things better by collecting it, connecting it to other data, and analyzing it using algorithms. If the governance fails, the models fail too. Zuboff (2019) says that this change has created a new economic logic in which data collected from users' behaviors is used not only to make products better but also to build prediction markets based on what people are likely to do in the future. No matter how you look at this larger thesis, it does capture an important operational reality: companies that used to see data governance as a defensive, cost-minimizing function now see it as a necessary condition for AI-driven value creation, since models trained on data that was not managed well pick up on its gaps, biases, and inconsistencies.

This reframe has effects on the organization. Chief data officers used to only be concerned with regulatory reporting, but now they are seen more as strategic partners to the business and are in charge of data quality systems that directly support AI development pipelines. Data governance councils used to meet to approve retention schedules. Now, they look at the history of training datasets that will be used for models that interact with customers. The field has not given up on its compliance roots, but it has grown to include data quality engineering, metadata management, and more and more, reviewing the intended uses of data in an ethical way.

2.2 The Data Lifecycle as a Governance Unit

Another change that is linked is the move away from using individual datasets as the unit of governance and toward using the whole data lifecycle as the unit of analysis. The lifecycle view follows data from when it is collected or created to when it is stored, merged, changed, used in analytics or model training, and finally when it is deleted or archived. At each stage, different governance needs come up. Collecting data brings up issues of consent, purpose limitation, and data minimization; storing data brings up issues of security and access control; changing data brings up issues of provenance and reproducibility; and using data in AI systems brings up issues of fairness, representativeness, and explainability.

According to the lifecycle view, responsible AI is especially important because many of the problems with algorithmic systems start much earlier, when decisions are made about what data to collect, how to label it, and which groups it represents. Instead of problems with the learning algorithm itself, Barocas and Selbst (2016) showed that unfair results in automated decision systems are often caused by biased training data, outcome variables that are labeled incorrectly, or proxies that are linked to protected traits. This finding is important for designing governance because it means that interventions that only focus on training or evaluating models will miss a lot of the risk. This is because the dice are often cast long before any modeling code is written. For responsible AI to work, data governance needs to include awareness and control at the point of data collection and curation, not just at the point of deployment.

Lifecycle governance also means that an organization is responsible for different things. Instead of putting all the responsibility for data quality on one central function, mature governance frameworks spread it among data stewards in business units who understand how data is generated while still having a central point of control for cross-cutting standards, audits, and problems that need to be escalated. Large financial institutions and tech companies now use this federated model a lot. It tries to balance the need for domain-specific judgment with the need for consistent, auditable standards across the whole organization.

Metadata management has become an important part of lifecycle governance, especially for organizations that have to follow the rules for documentation set by laws like the EU AI Act. With detailed metadata that includes where the data came from, how it was collected, any known limitations, and the basis for consent, an organization can find out, often years after a dataset was first put together, if a certain data source is still legally and morally acceptable for a new use. A common problem is that there is not enough metadata. Many companies find out that they can not find the original consent or where the datasets came from after years of copying, merging, and using them in different ways across the company. They only find out this when they do a data protection or AI compliance review. So, investing in metadata infrastructure is not just a nice-to-have for technical reasons; it is also becoming a more important part of being able to defend yourself in court.

It is also important to see data quality as a concept with many parts, not just one trait. Each of these things—completeness, accuracy, timeliness, consistency across systems, and representativeness of the population a model is meant to serve—is important on its own, but none of them are enough by themselves. A dataset might be very accurate and complete for the population it was collected from, but not for a model that is meant to serve a larger or different population. This is something that technical data-quality tools that only check for completeness and internal consistency do not show. This is one of the main reasons why lifecycle governance frameworks need a clear statement of the intended use and population along with any dataset that is made available for training the model. This way, the dataset's representativeness can be checked against a clear target instead of being left open-ended.

3. Foundations of Responsible AI

3.1 Principlism and Its Limits

Over the past ten years, a wave of ethical guidelines from governments, industry groups, and international organizations has had a big impact on the conversation about responsible AI. Jobin, Ienca, and Vayena (2019) looked at 84 of these kinds of papers and found that they all agreed on five principles: honesty, fairness and justice, not hurting others, taking responsibility, and privacy. The AI4People framework, which is related and important, suggested adding a fifth principle of explicability, which includes both intelligibility and accountability, to the four classical bioethics principles of beneficence, non-maleficence, autonomy, and justice (Floridi et al., 2018). In response, the Organization for Economic Co-operation and Development (OECD) published a

similar set of values in 2019 in its Recommendation on Artificial Intelligence. These values include inclusive growth, human-centered values, transparency, robustness, and accountability. Dozens of member and non-member states have since adopted or referenced this document (OECD, 2019).

While there has been convergence at the level of abstract principles, there has not been convergence at the level of implementation. Jobin et al. (2019) said that there were big differences in how the five core principles were known, who and what they were thought to apply to, and how they should be put into practice. Mittelstadt (2019) gave a stronger criticism, saying that the comparison between principled writing and medical ethics is not perfect: medicine has a history of professional culture, licensing, and liability structures that make principles work in real life, but the AI industry does not have any similar systems that hold professionals accountable. Mittelstadt said that without these kinds of checks and balances, ethical principles might just act as a way for companies to protect their reputations by saying they will be responsible without changing the incentives or ways things are made.

Empirical study into corporate AI ethics programs has repeatedly shown that stated principles do not always match up with how engineering teams work when they are under normal product deadlines and performance metrics. This supports the criticism. It is not likely that ethics statements will have much of an effect on day-to-day development decisions if they are not turned into specific design requirements, testing protocols, or escalation procedures. Organizations that want to make responsible AI work should learn that principles are a good place to start, but they are not enough to guide AI on their own; they need to be built into systems with consequences for breaking the rules.

One of the five convergent principles that Jobin et al. (2019) identified is privacy. This principle shows how different people still interpret things even when there seems to be agreement. Solove (2006) showed that privacy is not just one concern, but a group of related ones. These include protecting against information collection, putting together of previously separate data points, unauthorized disclosure, and prying into personal decision-making. Each of these involves different harms and needs a different set of responses from government. An AI ethics statement that promises to protect privacy but does not say which aspect of privacy is at risk in a specific system makes it hard to know how that promise should affect actual design decisions. Solove's aggregation-based privacy harm is increased when a recommendation system uses otherwise harmless behavioral signals to infer a sensitive trait, like a health condition or sexual orientation. This happens even when no individual data point disclosed would be considered sensitive. A general privacy principle probably would not be able to tell the difference between these situations without more detailed taxonomic guidance, which is what Solove's framework gives.

3.2 From Principles to Practice: Accountability Mechanisms

Because principlism has its flaws, more and more research and practice is focusing on concrete accountability methods that can turn high-level commitments into observable organizational behavior. According to Raji et al. (2020), there should be an end-to-end internal audit framework

for AI systems. This framework should include scoping documents during the design stage, testing to see if performance is different across subpopulations before deployment, and monitoring after deployment for model drift and new harms. Their framework does not see accountability as a single review gate, but as an ongoing process that is part of the normal software development lifecycle. This is similar to the way that data governance is moving toward a lifecycle orientation. In a related area of study, the sociotechnical aspects of algorithmic fairness have been looked at. Self, Boyd, Friedler, Venkatasubramanian, and Vertesi (2019) said that a lot of technical efforts to make things more fair do not work because they look at fairness as something that can be checked in a model, without thinking about how the system will work in real life. They found several common ways that systems fail. These include thinking that a fairness metric that works in one deployment situation will work the same way in another, and seeing the technology as a separate thing instead of as part of a bigger sociotechnical process that includes people making decisions, feedback loops, and different ideas of what is fair. In addition to technical fairness testing, their work suggests that responsible AI governance should include subject expertise and feedback from the affected community. Fairness should not just be seen as a mathematical property that needs to be optimized.

Another way to hold people accountable is to make sure that automated decisions can be explained and contested. Wachter and Mittelstadt (2019) looked at how current data protection law deals with what they called "high-risk inferences." These are predictions or classifications made from data using statistical or machine-learning methods that individuals cannot check or dispute. They suggested adding a right to reasonable inferences to the existing rights of access and rectification. This proposal is an example of a larger idea in the literature about accountability: as AI systems make decisions based on inferred traits rather than directly observed traits, they need accountability mechanisms that go beyond the usual focus on the accuracy of input data and look at the inferential process itself.

There is another difference between algorithmic accountability and organizational accountability. Algorithmic accountability is the technical ability to explain or audit a model's outputs. Organizational accountability is the bigger question of who or what in an organization is responsible for the design, deployment, and effects of a system, and how that responsibility can be enforced. Cheney-Lippold's (2017) study of algorithmic categorization shows why this difference is important: many of the bad things that happen with automated systems are not caused by a single wrong prediction, but by the effect of constantly changing, opaque categorizations that affect the opportunities, prices, and content that a person sees over time. This kind of harm is hard to catch by accountability systems that only focus on challenging individual choices. For this reason, good governance needs organizational accountability structures that can keep an eye on the overall, long-term effects of a system in use, not just ways to dispute individual outputs after the fact.

4. The Legal and Regulatory Architecture

4.1 The GDPR as a Data Governance Baseline

The General Data Protection Regulation (GDPR) of the European Union, which has been in effect since 2018, is still the most important law affecting how data is managed around the world (European Union, 2016). Its importance for responsible AI lies not so much in a single provision as in the governance infrastructure it requires: companies that process personal data must keep records of their activities, do data protection impact assessments for high-risk processing, hire data protection officers when certain thresholds are met, and protect data by design and by default. These rules have made many of the lifetime governance practices talked about in Section 2 official. They are now required by law instead of being up to individuals to choose to follow.

The GDPR's rules on automated decision-making are especially important for AI that is responsible. Article 22 says that people have the right not to be forced to follow a decision that was made only by computers if it has legal or similarly important effects. However, there are certain exceptions and protections that apply, such as the right to get a human to help and the right to challenge the decision. Kaminski (2019) says that the GDPR's approach is like a type of binary governance because it combines individual rights, like the right to be informed, with systemic, collaborative governance duties, like doing impact assessments and talking to supervisory authorities first. She said that this hybrid model is better seen as a pair that works well together than as a choice between individual empowerment and systemic oversight. This is because individual rights alone are not good at finding harms that happen to a lot of people, and systemic oversight alone might not be able to protect the interests of any one person who is affected.

Another reason why the GDPR has such a big effect on global data control is that it can be used in other countries. The rule applies to any organization in the European Union that handles personal data of people living in the EU, no matter where that organization is based, as long as the processing is necessary to provide goods or services to those people or keep an eye on how they behave. This international reach, along with the size of the European market, has caused many multinational companies to adopt GDPR-aligned data governance practices as a global standard instead of keeping separate regimes for each jurisdiction. This is called the "Brussels Effect," and it is talked about in more detail in Section 7.2.

As a governance tool, the data protection impact assessment is especially important because it makes many of the lifecycle principles we talked about in Section 2 work for each individual project. When processing is likely to put people's rights and freedoms at high risk, the organization must describe the processing operation, evaluate its necessity and proportionality in relation to its stated purpose, list the risks to the people who will be affected, and keep track of the steps it took to lower those risks, all before the processing takes place. This requirement has been interpreted more and more to include a look at where the training data came from and how representative it is, how the model is expected to behave across relevant subpopulations, and the ability for humans to review any bad decisions made by the system. This effectively turns the impact assessment from a privacy tool into a broader tool for anticipatory AI governance, even though the GDPR was created several years before the current generation of large-scale machine-learning systems.

4.2 The EU AI Act and Risk-Based Regulation

Where the GDPR governs the processing of personal data generally, the European Union's Artificial Intelligence Act, which came into force in 2024, creates a dedicated regulatory framework specifically for AI systems, organized around a risk-based tier structure (European Union, 2024). The Act only allows a few AI practices that are thought to be too dangerous. These include some types of subliminal influence and social scoring by the government. It requires a wider range of high-risk AI systems, such as those used in hiring, credit scoring, law enforcement, and critical infrastructure, to follow a set of rules that include risk management systems, data governance and quality requirements for training datasets, technical documentation, human oversight, and monitoring after the market has been released. For example, chatbots and other systems that do not pose a lot of risk are only required to be transparent in a limited way. The Act does not regulate these systems very much.

Writing during the legislative negotiations of the Act's precursor draft, Veale and Zuiderveen Borgesius (2021) said that the risk-tier approach is very different from the GDPR's more uniform, rights-based model. They said that this makes the regulatory center of gravity shift from ex post individual rights of redress to ex ante conformity assessment and standardization. They said that this method requires a lot of interpretation based on putting a system into the right risk tier, which is not always easy to do when a system has more than one use or is later put to higher-risk uses than it was originally meant for. This problem with classification has direct effects on data governance because a system's data quality and documentation duties depend on which tier it is in. This means that companies that use general-purpose AI models in many downstream applications need to keep governance systems that can keep track of and reclassify use cases as they change.

The Act's data governance rules for high-risk systems are unique because they make a clear legal link between bad data and the chance of unfair or unreliable results. They say that training, validation, and testing datasets must be relevant, sufficiently representative, and, as much as possible, error-free. They also need to take into account the features of the specific location, context, or function where the system is meant to be used. In this rule, the idea that data quality before model training is a major factor in determining harm later on is made official by the law. This idea comes from the fairness and accountability literature that we talked about in Section 3.

The Act's treatment of general-purpose AI models, such as large foundation models that can be changed to do a lot of different jobs later on, makes governance even more complicated. Because a single general-purpose model can be used in downstream applications spanning all of the Act's risk tiers, from low-risk consumer applications to high-risk uses in employment or credit decisions, the Act requires basic transparency and documentation from providers of such models no matter what downstream use they are put into. The downstream deployer that puts the model into a specific application is responsible for risk-tier-specific obligations. As responsibility is spread out along the AI value chain, providers and deployers must keep up contractual and technical ways to

pass important information, like model cards that describe the characteristics and known limits of training data, from the point of model development to the point of application-specific deployment. This is because a deployer can not meet its own data governance obligations under the Act without being able to see the decisions made by a model provider it may never directly interact with.

4.3 Cross-Border Data Flows and Regulatory Fragmentation

Although the GDPR and the AI Act have had a big impact outside of the European Union, they work with a growing number of different national and regional data protection and AI laws. These include sector-specific US frameworks, China's data security and algorithm laws, and comprehensive laws in places like Brazil, South Korea, and India. Because there are so many of them, organizations that work in more than one jurisdiction have to deal with different, and sometimes conflicting, rules about data localization, cross-border transfer mechanisms, consent standards, and algorithmic transparency duties.

As a result, data governance needs to include jurisdiction-aware data architecture. This means that data needs to be classified not only by how sensitive it is and what it is used for, but also by the regulatory regimes it is subject to. It also needs to have transfer mechanisms, like standard contractual clauses or adequacy determinations, that are actively maintained instead of just being adopted once and forgotten. For AI governance in particular, this fragmentation makes things more complicated because one model may be trained on data from several jurisdictions, each with its own rules on how representative the data must be, how long it must be kept, and how it can be used. This means that organizations have to keep track of not only what data was used, but also the legal grounds under which it was collected and processed in each jurisdiction.

Data localization rules say that certain types of data must be kept or processed within the borders of a certain country. These rules add another architectural limitation that affects AI development in a unique way. When training a model, it is often best to use as many relevant examples as possible. However, localization requirements that split up a global dataset into storage environments that are legally separate from each other can have a big impact on the statistical power and representativeness of the model that is produced. This is especially true for applications that serve small national markets and can not create a large enough or diverse enough training corpus on their own. In response, businesses have come up with a variety of technical and organizational strategies. These include federated learning methods that train models on distributed, legally separate data without sending the raw data across borders, and creating synthetic data to mimic the statistical properties of a limited dataset in a way that makes it easier for everyone to use. Each of these strategies has its own governance implications. For example, federated learning raises questions about how easily a training process can be inspected when it is spread across multiple legal and organizational boundaries. Similarly, synthetic data raises questions about how well the synthetic distribution maintains the properties, including any biases that may be built into it, of the original data it is meant to approximate.

5. Business Strategy and Organizational Implementation

5.1 Governance Structures: Roles, Committees, and Culture

Putting the above-mentioned legal and moral requirements into practice in an organization requires careful structural design. Companies that have mature responsible AI programs often set up a cross-functional AI governance committee. This committee should look over high-risk use cases before they are deployed and then on a regular basis afterward. It should be made up of people from legal, data science, information security, product, and, more and more, roles that are specifically dedicated to AI ethics or responsible AI. This committee structure is similar to the data governance council model described in Section 2. In many organizations, it is also integrated with that model. This shows that data and AI governance should be seen as a single, integrated role, rather than two separate ones.

A big help in organizing these business processes is the AI Risk Management Framework from the National Institute of Standards and Technology (NIST, 2023) for groups working in the US or looking for a voluntary standard that is recognized around the world. AI risk management is based on four main tasks, which are govern, map, measure, and manage. The govern function looks at the cultural and structural foundations of responsible AI, such as policies, accountability structures, and the diversity of the people working on AI development teams. The map function finds and groups risks that are unique to a system and how it is used. The measure function rates risks on a scale of one to ten, and the manage function decides which risks to focus on and how to handle them. Because the framework is not specific to any one industry and is voluntary, it has been used as a common language by organizations that have to meet both mandatory requirements under laws like the EU AI Act and voluntary obligations to stakeholders in places where there are not any similar mandatory laws.

In addition to formal structures, organizational culture is known to affect whether governance tools work as planned or just serve as a show of compliance. When engineering teams see responsible AI review as an outside force that comes in late in the development process instead of as an important part of the design process, they tend to see it as a problem that needs to be fixed instead of a source of design input. When an organization has deeper integration, governance and ethics experts are usually involved from the very beginning of problem-solving and dataset selection. This is in line with the lifecycle governance model we talked about in Section 2.2. At least some of the technical performance evaluation is tied to responsible AI outcomes like fairness testing completion or documentation quality.

5.2 Responsible AI as Competitive Differentiation

In the research on digital trust in management, one theme that comes up a lot is that responsible data and AI practices can be both defensive and offensive. This is because they can help a company stand out from others in the same market, especially when there are many providers offering mostly the same functions. More and more, enterprise customers in regulated industries use AI governance maturity as a part of their vendor due diligence processes. They see documented data lineage, bias

testing protocols, and model documentation as requirements for buying rather than nice-to-have features. In this way, ethical and regulatory expectations are successfully passed down the supply chain. This makes companies that are not directly affected by laws like the EU AI Act more likely to adopt similar internal standards in order to keep doing business with regulated customers.

This new way of thinking about strategy does not get rid of the tension between the short-term costs of investing in governance and the longer-term, harder-to-quantify benefits of trust and avoiding risk. This tension is especially strong for smaller organizations that do not have the resources of large incumbents to build dedicated governance functions. But it does show that looking at responsible AI only as a cost center tends to undervalue its importance in getting into new markets, keeping customers, and avoiding costly fixes after deployment, fines from regulators, or damage to the company's image after a public failure.

5.3 Measuring What Matters: Metrics and Audits

One problem that keeps coming up with implementation is how to measure responsible AI performance in a way that is specific enough to help management make decisions, as opposed to the more general principles we talked about in Section 3.1. Organizations have agreed on a mix of process metrics, like the percentage of high-risk AI use cases that had a documented impact assessment done before they were put into action, and outcome metrics, like the difference in performance between demographic subgroups on relevant tasks. Raji et al. (2020) said that audit trails that show why design decisions were made, the datasets and evaluation protocols that were used, and the signatures that were obtained at each stage are just as important as any single quantitative fairness metric. This is because they allow for accountability after the fact if a system has an unexpected or harmful outcome after deployment, which is something that aggregate performance metrics alone do not allow for.

More and more, external and independent audit is being talked about as a supplement to internal measures. This is similar to financial audit, but the field does not have the standard methods, licensing, and liability rules that give financial audit its weight as evidence. More generally, Pasquale (2015) said that the lack of transparency in private algorithmic systems, which is caused by trade secrets and technical difficulties, makes it hard for outsiders to hold companies accountable. This is an issue that the EU AI Act's conformity assessment and documentation requirements and voluntary third-party audit initiatives try to solve, but with varying levels of success.

5.4 Talent, Skills, and Organizational Learning

A less talked about but important factor in how well governance works is the availability of staff who are technically proficient in data science and machine learning and also have a basic understanding of the relevant legal and moral frameworks. Purely technical hires, even if they are good at building models, often do not know when a design decision raises a data protection or fairness concern that needs to be brought to the attention of governance for review. On the other hand, purely legal or compliance hires often do not have enough technical knowledge to tell if a

proposed technical mitigation really addresses a known risk or just makes it look like it does. Organizations that have good governance functions usually put money into cross-training, making sure that staff with a mix of technical and legal knowledge work in both engineering teams and central governance functions. They also treat this mix of skills as a separate professional competency that needs to be hired for and developed, rather than assuming that it will happen naturally as specialists from different fields work together.

Structured post-incident review after any governance failure or near-miss, as well as regular retrospective analysis of previously approved use cases in light of subsequent outcomes, are two more ways that mature governance functions can be told apart from those that are still just procedural. Because AI systems and the situations in which they are used change after they are first approved, a governance function that sees approval as a one-time event rather than the start of a long-term relationship of monitoring will likely let risk build up over time, even if the review processes were thorough at the start. So, the lifecycle approach that is pushed throughout this piece includes the institutional learning of the governance function as well, not just the technical systems that it is in charge of.

6. Sectoral Illustrations

6.1 Financial Services

Because model risk management rules came before the current wave of AI-specific rules, the financial services industry is one of the most mature places to see how data governance and responsible AI work together. Credit scoring and underwriting models have been closely looked at for decades as part of fair lending. This existing regulatory framework has made it easy to add AI-specific fairness testing, which requires lenders to show that a model does not unfairly affect protected groups even when no variable directly encoding a protected characteristic is used as an input. This is more likely to happen because machine-learning models can create proxies for such characteristics from data that seems neutral (Barocas & Selbst, 2016). Creditworthiness assessment is clearly labeled as a high-risk use case under the EU AI Act. This means that institutions working in the European Union have to follow all of the data control, documentation, and human-oversight rules set out in Section 4.2.

Fraud detection is a related but separate governance challenge because the operational value of these systems depends on being able to adapt to changing fraud patterns all the time. This creates tension with governance processes that are based on reviewing things at set times. As a result, institutions have created continuous monitoring pipelines that keep an eye on model drift and flag statistically significant changes in decision patterns so that governance can be reviewed more quickly. This is in line with the lifecycle governance model pushed for in this article, where oversight is built into a system from the beginning and continues throughout its operational life.

Insurance underwriting is another example because insurers have used actuarial models to price risk for a long time—more than a hundred years before AI. However, these models are now being replaced by machine-learning models that are trained on much larger and more detailed data sets,

such as behavioral and telematics data that traditional actuarial methods can not access. As we know from fair lending, this expansion brings up the worry that seemingly neutral behavioral variables may be used as stand-ins for protected traits like age, disability, or socioeconomic status. This is especially true when the underlying data shows that people have had different levels of access to affordable insurance in the past. Several places have responded by adding to the insurance rate-filing and anti-discrimination review processes that require insurers to document and explain the variables used in AI-based pricing models. This is an example of adding AI-specific data governance requirements to a sectoral regulatory infrastructure that was already in place. This is a common pattern in financial services and other sectors with mature pre-AI regulatory infrastructure may find it useful to use as a model.

6.2 Healthcare

When AI is used in healthcare, it brings up very important questions about data governance. This is because clinical data is very sensitive, decisions about diagnosis and treatment have a lot at stake, and there is a well-known risk that clinical datasets show unfair access to care in the past, which can be stored in predictive models trained on such data. The EU AI Act says that AI systems that are meant to be used as a safety feature of a medical device or as a medical device themselves are considered to be high risk. This means that they need to follow extra rules about how to be governed on top of the rules that already apply to medical devices.

One problem that comes up a lot in this area of governance is making sure that the training data is representative of all demographic groups. This is because clinical trial populations and electronic health record systems have historically underrepresented some groups. This is a problem that affects any predictive model that is trained on this data. To solve this problem, data governance needs to step in before the model is built. This includes planning how to collect different kinds of data and writing down any known gaps in the dataset's coverage. This is in line with the EU AI Act's requirement that training data should be sufficiently representative of the people and situations where the system will be used. Healthcare also shows the limits of purely technical fairness interventions emphasized by Selbst et al. (2019). This is because a model that is tuned to work equally well across demographic groups on a training dataset may still produce unfair results if it is used in a clinical workflow that gives different groups different access to the interventions the model suggests.

The clinical setting also shows how important human oversight design is in a way that formal availability of a human reviewer does not. When doctors use an AI-generated risk score or diagnostic suggestion during a rushed consultation, they might trust the system's output more than a formal human-in-the-loop requirement would allow. This is known as automation bias, and it happens especially when the system's outputs look very accurate without actually being accurate based on the model's reliability in that specific clinical setting. For this reason, good governance in this situation requires not only checking if a human reviewer is officially allowed to override a system's output, which many deployed systems do, but also checking if the interface design,

workflow integration, and clinician training actually allow for meaningful, informed override in practice. This is a difference that directly affects how the human-oversight requirements in laws like the EU AI Act should be interpreted and checked in clinical deployments in particular.

6.3 Public Sector

Using AI in the public sector, from finding fraud in benefits administration to predictive policing and automated decision support in immigration processing, comes with its own set of governance challenges. These come from the fact that the government has the power to force people to do what it wants, as well as constitutional and administrative law limits on how it can make decisions. The EU AI Act lists as "high risk" a number of public-sector use cases, such as systems that check if someone is eligible for essential public assistance benefits and services. This is because policymakers believe that algorithmic mistakes in these areas have especially bad effects on the people affected, who often do not have many other options.

Government agencies also have to deal with a part of due process that is not as important in the private sector: in many places, administrative law says that decisions that affect people's rights must come with a reasoned explanation that can be used for judicial or administrative review. This requirement is similar to the explicability principle we talked about in Section 3.1, but it is a legal requirement in the public sector instead of just an ethical goal. Several places have made it so that any automated decision that could hurt a person must be reviewed by a person before it can be legally used. This makes the human oversight requirements that the EU AI Act sets at the system design level work in real life cases.

Procurement practice is another unique tool that can be used to govern the public sector that does not have many close equivalents in the private sector. Because government agencies usually buy AI systems from outside companies instead of building them themselves, and because public procurement is subject to strict legal requirements around openness and fair competition, some places have started to include AI-specific governance requirements directly into procurement rules. For example, as a condition of contract award, vendors may have to disclose the characteristics of training data, submit to independent bias testing, or agree to ongoing monitoring obligations. This way of governing through procurement has the big benefit of using leverage before a contract is signed, when the government still has a lot of negotiating power. This is in contrast to only using ex post regulatory enforcement against a vendor whose system has already been used by multiple agencies. It also provides a way for public-sector governance requirements to indirectly change vendor practices that then show up in that vendor's private-sector offerings as well.

7. Tensions and Unresolved Challenges

7.1 Innovation versus Precaution

A persistent issue in both academic and policy debate is the right balance between precautionary regulation meant to stop harm from AI and the risk that too strict governance requirements will put regulated organizations at a disadvantage compared to competitors operating under lighter-

touch regimes, or will be impossible for smaller organizations that do not have the compliance resources of large incumbents. Some people who are against risk-based frameworks like the EU AI Act say that classification thresholds make it easier to engineer around high-risk labels instead of directly addressing the risks. On the other hand, people who support them say that the alternative—either no sector-specific AI regulations or a single standard that applies to all applications, would either not address real harms or put too many costs on low-risk ones. This is a real empirical and policy question that reasonable analysts do not agree on, and there is not a lot of evidence to back it up yet because AI-specific regulations just started going into effect.

A related part of this debate is how the costs of compliance affect businesses of different sizes. Large companies that are already in the market are usually better able to handle the fixed costs of setting up dedicated governance functions, hiring specialized outside counsel, and keeping up with the documentation infrastructure that risk-based frameworks need. On the other hand, smaller companies and startups may have to deal with compliance burdens that are disproportionate to their revenue, which could strengthen the market position of already-dominant companies even if that is not what the regulation was meant to do. Some places have tried to solve this problem by creating regulatory sandboxes, which let smaller, less-resourced businesses test new AI applications while being closely watched by regulators and with fewer formal compliance duties for a set trial period. These sandboxes also make it easier for businesses of all sizes to follow the rules. It is still not clear if these changes are enough to counteract the scale savings of compliance infrastructure. This is similar to the tradeoff between innovation and caution in general, and it needs more research as the relevant frameworks get better.

7.2 Global Fragmentation and the Brussels Effect

Some people think that the Brussels Effect, which is the extraterritorial effect of European data protection and AI regulations, will lead to some level of de facto global harmonization as multinational companies use European standards as their operational baseline instead of keeping their own systems that are specific to their own jurisdictions. Others are less sure, pointing out that the US, China, and other major countries have developed different, and sometimes competing, regulatory philosophies. These philosophies are based on different ideas about how to balance innovation, individual rights, and state control over information flows. Whether there will be more convergence or continued fragmentation is still an open empirical question. In the meantime, global organizations must design governance systems that can handle both scenarios and meet the strictest requirements in each area of operation, without assuming that regulatory convergence will make this easier in the near future.

7.3 The Accountability Gap

Lastly, a lot of research, like Pasquale (2015) on algorithmic opacity and Raji et al. (2020) on internal audit, agrees that it is hard to hold AI systems externally accountable, even when there are strong internal governance processes in place. This is because people who are affected by the systems, researchers, and regulators usually can not get their hands on the technical

documentation, training data, and decision logs that are needed to independently check an organization's claims about how fair or safe its systems are. The EU AI Act's documentation and conformity assessment requirements are a big step in the right direction, but how well they work in real life will depend on how strong and independent the notified bodies and market surveillance authorities are. At the time of writing, these institutions are still being built up across member states. As some of the scholars this article talks about have pointed out, closing this accountability gap may require more than what is currently required by law in most places. For example, algorithmic impact assessments that are required and can be audited by a third party may be needed, as well as rights of access for people who are affected and their representatives.

7.4 Measurement and Standardization Gaps

One last problem that has not been solved is that there are not any standard, proven metrics for many of the things that responsible AI frameworks ask companies to check, like how fair, robust, and explainable the AI is. When it comes to financial auditing, generally accepted accounting principles have been honed over more than a hundred years of professional practice and legal precedent. But when it comes to algorithmic fairness, there are multiple, and sometimes mathematically incompatible, formal definitions. This means that a model that meets one recognized fairness criterion may necessarily violate another when it comes to the same protected groups and the same underlying data. This is not just a technical problem; it means that a claim of fairness compliance only makes sense in a certain concept and setting. This subtlety is easy to miss when claims of fairness compliance are made in broad terms to non-technical stakeholders, regulators, or the public. The EU AI Act and other standard-setting bodies have started to close this gap by referencing technical standards. However, there are a lot of different scientific and mathematical definitions of fairness, so standardization can only spell out how to choose and defend a definition that fits the situation. It can not solve the plurality itself. So, organizations should not see any one fairness number as a complete sign of ethical compliance. Instead, they should see it as a partial and situation-specific test.

8. Toward an Integrated Framework

Putting all of this analysis together, we suggest that companies that want to use responsible AI should see data governance, algorithmic accountability, and business strategy as three closely connected parts of a single governance architecture, rather than three separate functions that are only loosely coordinated. Based on the analysis in this article, four design concepts can be drawn. First, governance should not be based on single reviews at certain points in time, but on the whole lifecycle of the data and models. As shown in Sections 2 and 3, many of the biggest risks in AI systems come from choices about collecting data and sanitizing it made long before the model is trained. These risks also change after the system is put into use because the model drifts and the use contexts change. Lifecycle governance needs to be able to see and control everything from collection to removal. This can not be done with just one pre-deployment sign-off.

Second, following the law should be seen as a floor, not a roof. The GDPR, the EU AI Act, and similar laws in other countries set minimum requirements that are based on what is politically and administratively possible, not on the risk profile or stakeholder expectations of any one organization. If organizations use these thresholds as a stopgap rather than as a starting point for more context-specific risk management, they will consistently underinvest in governance compared to the actual risk they face, especially for new uses that are not yet covered by existing regulatory categories.

Third, accountability systems need to have both internal auditing tools and real ways for outsiders to look at them and make complaints. As we talked about in Section 7.3, internal governance processes, no matter how well they are designed, can not fully replace external verification. If a company wants to keep the trust of its stakeholders over time, it should invest in documentation, audit trails, and contestability mechanisms that would still be reliable under independent, adversarial review, instead of mechanisms that are only meant to please internal or self-selected external auditors.

Fourth, investments in governance should be seen and talked about within the company as something that adds value to the company, not just something that costs money. As we talked about in Section 5.2, this new way of looking at things is backed up by the fact that governance maturity is becoming more and more of a requirement for entry into business-to-business and regulated markets, and by the fact that governance failures have well-known financial, operational, and reputational costs that show up after deployment. If you only think of governance as a set of rules that are put on the product and engineering teams from the outside, you are more likely to get the kind of symbolic, poorly integrated compliance behavior that the critique of principlism in Section 3.1 says is the main reason why corporate AI ethics initiatives have failed so far.

These four principles work hand-in-hand rather than separately, and companies that want to follow them do not have to or usually can not do so across their whole business at the same time. A common and usually good way to do things is to start with a small group of high-profile, high-risk use cases. These are chosen because failing to govern these cases would have the biggest legal or reputational impact. Then, the lifecycle governance processes, documentation practices, and cross-functional review structures that were created for these initial cases are used as a model for later, lower-risk applications across the organization. Before trying to apply governance processes to all AI applications, this staged approach lets an organization build institutional capability and organizational learning, as we talked about in Section 5.4. It also tends to produce more durable governance capability than trying to impose a uniform, fully specified governance framework across the whole organization at once, before any part of the organization has had a chance to work within it.

9. Conclusion

Responsible artificial intelligence can not be made by just one field working on it. No matter how complex the technical fairness interventions are, they can not make up for poorly managed training

data or deployment contexts that were never looked at by someone with the right subject knowledge. Legal compliance, no matter how thorough, sets minimum standards that are based on political feasibility rather than the actual risk profile of each company. And a business strategy that sees governance only as a cost that needs to be cut will usually lead to the very governance problems that the strategy was trying to avoid. This article makes the case that data governance, responsible AI practice, and business strategy should all be seen as three interdependent parts of a single organizational capability. They should not be seen as three separate functions that are only loosely coordinated at the edges. This way of thinking is necessary for AI systems to be effective, legal, and worthy of the trust that people and institutions put in them. As regulations change, especially as the EU AI Act moves from being signed into law to being fully implemented and as other countries create their own frameworks, this integrated approach provides a base that can adapt to new regulations and technological advances. This is because it is based on the structure of AI risk rather than a single, possibly temporary regulatory instrument.

Several of the questions raised in this article should be kept in mind as more implementation experience is gained. If longitudinal studies were done to see if companies that use the integrated governance architecture suggested here actually have fewer failures after deploying AI and if this investment in governance leads to measurable business outcomes like keeping customers or getting into new markets, it would make the case for seeing responsible AI as a source of business value rather than just a cost much stronger. Comparative research that looks at how the risk-based European model, the sectoral US approach, and other national frameworks work in practice, after each has built up a good record of enforcement and organizational adaptation, would also help with the ongoing debate over regulatory design that was talked about in Section 7. Until that evidence builds up, both organizations and policymakers have to work with real uncertainty. This article has tried to accurately describe this situation rather than prematurely resolve it.

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

DAMA International. (2017). DAMA-DMBOK: Data management body of knowledge (2nd ed.). Technics Publications.

European Union. (2016). Regulation (EU) 2016/679 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 (General Data Protection Regulation). Official Journal of the European Union.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018).

AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Kaminski, M. E. (2019). Binary governance: Lessons from the GDPR's approach to algorithmic accountability. Southern California Law Review, 92(6), 1529–1616.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce.

Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68.

Solove, D. J. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154(3), 477–564.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), 97–112.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review, 2019(2), 494–620.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.