Accountability for Responsible AI Practices: 
Ethical Responsibilities of Senior Leadership

Alex John London, Ph.D.  
K&L Gates Professor of Ethics and Computational Technologies  
Carnegie Mellon University  
Pittsburgh, PA 15213  

Copyright Alex John London  
7 Feb 2024

Abstract: Despite the proliferation of guidance documents, the role and responsibilities of senior corporate leaders for ethical and responsible development, procurement and use of artificial intelligence (AI) remains under-examined. Instead, attention tends to focus on developers and other stakeholders with primarily technical expertise. This paper argues that senior leaders play a fundamental role in determining whether the development, procurement, and use of AI systems conforms to and satisfies the practices and requirements of ethical and responsible AI. First, there are multiple points across the development and use of AI systems where decisions must be made that require the integration of “technical” and “non-technical” considerations. Second, in some cases senior leadership are a stakeholder that directly influences such decisions. Third, senior leaders indirectly influence these decisions because of their special role in creating the norms, practices, structures, and division of labor within an organization. Together, these norms, practices, structures, and division of labor constitute the ecosystem within which AI systems are developed or procured and implemented. Finally, senior leaders have unique responsibility for two key pillars of ethical and responsible AI: governance and responsibility. Ensuring that senior leadership act to uphold the norms and requirements of ethical and responsible AI is critical to averting and closing loopholes in governance.

Disclosure: This paper was supported by a grant from the SWIFT Institute.

Introduction

It may strike some as paradoxical that the success of artificial intelligence (AI) systems hinges crucially on the responsible and ethical organization of human behavior. This air of paradox derives, in part, from the appearance of conflict between this claim and the very nature of AI, understood loosely as computational systems capable of emulating human intelligence to the point that they can solve problems that were, prior to the development of these systems,
the sole province of human intelligence.\(^1\) It also derives, in part, from a tension with the rationale for developing and deploying AI systems in most enterprises, which is to achieve a cluster of benefits that ultimately derive from speed and scale that likely could not be achieved by human decision makers alone.\(^2\) Nevertheless, AI systems do not produce themselves. During development, AI systems depend on a division of labor that often cuts across different parts of an organization, sometimes reaching across different organizations, and potentially involves entities in different sectors (such as the private and public sectors). Similarly, on their own, AI systems often do not produce tangible benefit. They are, instead, the most visible element within a much larger ensemble of knowledge, practices, and processes that constitute a socio-technical system.\(^3\) The ability of this larger system to achieve in actual practice the benefits the AI is designed to produce thus hinges on an organization’s ability to integrate these complex computational systems into a larger workflow within a larger organizational context.

Understanding the dependence of AI systems on organizational and human behavior has three key benefits. First, it makes salient an aspect of ethical and responsible AI that remains underdeveloped. In particular, recent revelations of unmet expectations,\(^4\) \(^5\) \(^6\) unfair practices that exacerbate social inequalities\(^7\) \(^8\) and histories of oppression and exclusion,\(^9\) \(^10\) or harmful

---

products\(^{11}\) have driven the proliferation of frameworks to promote ethical or responsible AI.\(^{12}\) These frameworks often focus on norms, principles, or rules that are requirements to which AI systems should aspire, or constraints that they should not violate.\(^{13}^{14}\) Typically, frameworks for ethical AI appeal to values such as respect for human autonomy, prevention of harm, fairness, and explicability while frameworks for responsible AI appeal to a series of more specific requirements including: ensuring human oversight, robustness and safety, privacy and data governance, transparency, non-discrimination and fairness, societal and environmental wellbeing, and accountability.\(^{15}\)

Although such guidance is extremely important, stakeholders often struggle to understand how to implement these requirements in actual practice. This problem is particularly acute when it comes to senior organizational leaders since they may feel that, without a technical background in AI or ethics, they lack the skillset or knowledge base to take responsibility for decisions about the development, testing, or deployment of AI systems. Less well-developed, in other words, is an operational bridge from abstract principles to the way these requirements might be met through a division of organizational labor that involves a wide range of stakeholders with different kinds of expertise.

Second, the requirements for ethical and responsible AI often focus on developers or provide more concrete guidance to developers than to other stakeholders who, if they are mentioned at all, are often mentioned in passing.\(^{16}^{17}\) Similarly, discussions of specific values or

\(^{16}\) As an example, the otherwise excellent guidance from NIST mentions C-suite executives in Figure 3, but their role in responsible AI is not discussed in the document. National Institute for Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://doi.org/10.6028/NIST.AI.100-1
\(^{17}\) One notable exception is, World Economic Forum, Empowering AI Leadership: AI C-Suite Toolkit, Jan 2022 (https://www.weforum.org/publications/empowering-ai-leadership-ai-c-suite-toolkit/). This report provides a
requirements, such as fairness, often focus on mitigation strategies that can be applied at the stage of statistical modeling.\(^{18}\) Although developers are clearly an important stakeholder, and although statistical modeling is a critical aspect of AI system development, this emphasis obscures the role of other stakeholders and other choice-points across the ecosystem in which developers operate and in which AI systems are developed, procured, refined, deployed, and monitored. As one example, satisfying the requirements for ethical or responsible AI often requires time and resources. Allocating those resources can create apparent, and in some cases, genuine conflict with other goals or objectives of an organization. Approaches that do not address this potential for conflict, and which do not assign responsibility to senior leadership for resolving, mitigating, or managing it, likely permit loopholes in governance, understood as the ability of an agent avoid an ethical or regulatory requirement by altering the division of labor within an organization without changing the substantive ethical concerns associated with the underlying activity.\(^{19}\) Conversely, organizations may fail to generate both institutional and social value if they have the technical ability to produce and implement safe and effective AI systems, but senior leadership remains uncertain about whether or how to deploy these systems to realize these benefits at scale.

Third, ensuring the ethical and responsible development, procurement, and use of AI systems requires that stakeholders who are responsible for making decisions that impact this process are accountable for upholding appropriate norms. The failure to identify senior leaders as key decision makers in this domain creates the prospect of loopholes in governance—a situation in which one stakeholder is allowed to influence processes and outcomes without being accountable for abiding by relevant norms and principles. Such loopholes are pernicious for at least two reasons. First, they potentially allow some parties to act with impunity, averting


accountability for outcomes or actions that transgress relevant norms or principles. Second, to the extent that other parties are held accountable for practices or outcomes that result from these decisions, they create relationships in which blame is unfairly focused on parties who are not solely responsible for the relevant consequences.

These issues are of special relevance to banking and finance since these sectors have profound impacts on the lives of individuals, groups, social organizations and institutions and because regulatory guidance assigns ultimate responsibility for governance related to model risk to boards of directors and senior management. Such guidance reflects the underlying responsibility of senior leadership for ensuring that the human systems necessary for the ethical and responsible development and use of models, including AI systems, have been designed, implemented, resourced, and supported. This includes ensuring that relevant responsibilities have been distributed to appropriate decision makers, that standards, expectations, metrics, and other systems of accountability are in place, and, perhaps most importantly, that these roles, responsibilities, and requirements are backed by sufficient incentives, sanctions, or enforcement that they shape the behavior of key institutional actors.

**Senior leadership: a neglected focus for ethical and responsible AI**

Recognizing the role of senior leadership for ethical and responsible AI is hampered by a series of dynamics. The first involves the hype surrounding AI. Popular perceptions of AI include utopian claims about the ability of such systems to revolutionize social and business practices alongside dystopian visions of AI systems threatening the very existence of humanity. The sheer range of claims made about the promises and the perils of AI create an environment of uncertainty about its capabilities and feed a sense of **AI exceptionalism**, understood as the idea that AI is so uniquely different from other innovations or business tools that decisions about its use can only be made by parties with a special, technical expertise. AI exceptionalism is fueled, in turn, by a techno-focus in which AI is viewed as a complex computational system whose performance can be assessed in abstraction from the workflow of an organization and the social and business objectives of a firm. Together, these dynamics bolster the perception

---

that AI occupies a realm somewhat removed from the responsibilities of senior leadership. This is exacerbated by the extent to which discussions around the capabilities and limits of AI, as well as ethical discussions involving topics such as fairness or explainability, are often highly technical. When senior leadership do not themselves possess technical knowledge from computer science or statistics, they may regard these issues as falling under the expertise of specialists on whose judgment they must rely. This perception might be amplified by the proliferation of AI systems. A single firm might develop or procure AI systems to perform a wide range of tasks across diverse workflows (see box 1). For example, in its 2022 Annual Report, JP Morgan Chase reported that they have “more than 300 AI use cases in production today,” being carried out by “1,000 people involved in data management, more than 900 data scientists (AI and machine learning (ML) experts who create new models) and 600 ML engineers (who write code and put models into production).” The responsibility to manage these systems might seem to fall more naturally to unit managers who have a more detailed knowledge of the tasks for which these systems might be deployed and closer contact with technical experts developing or using them.

21 HTTPS://reports.jpmorganchase.com/investor-relations/2022/ar-ceo-letters.htm
Box 1: Sample applications of Artificial Intelligence (AI) in Banking and Finance

1. Investment management including customized portfolios
2. Forecasting cashflow, payments, consumer demand
3. High-frequency algorithmic trading
4. Personalized financial advice at scale
5. Credit underwriting, risk scoring
6. Anomaly detection / fraud detection
7. Assistance with regulatory compliance
8. Automated extraction of mortgage application information
9. Automated trade reconciliation

Understanding the role and responsibilities of senior leaders is facilitated by the recognition that AI systems are tools whose value is entirely dependent on whether they can effectively, efficiently, and equitably advance social and business objectives and whose success depends on whether they can be integrated into the workflow of the organization in ways that preserve and realize anticipated benefits. Within this context, four factors bring into sharper

---

22 General overviews include:


focus the critical role of senior leadership in ethical and responsible AI. It is helpful to first list these factors and then to discuss each in turn. First, there are multiple points across the development and use of AI systems where decisions must be made that require the integration of “technical” and “non-technical” considerations. Each of these considerations can have a direct bearing on the ethical and responsible use of AI. Second, in some cases senior leadership are a stakeholder that directly influences such decisions. Third, senior leadership can indirectly influence these decisions because of their special role in creating the norms, practices, structures, and division of labor that constitutes the ecosystem within which AI systems are developed or procured. Finally, senior leaders have unique responsibility for two key pillars of ethical and responsible AI: governance and responsibility.

**Non-Technical Aspects of AI Development: Constructing an AI-Development Portfolio**

How we describe the various stages in the lifecycle of AI development depends on what our goals are. For present purposes, our goal is to clarify the interplay between technical and non-technical considerations in the development of AI systems. It is important to emphasize that this boundary is not absolute. It can be fuzzy and difficult to draw, and the outline of this boundary may only emerge *during* the development process. In that sense, the goal of the present discussion is not to defend this distinction as clearer or more certain than it is, but to illustrate how the development of AI systems involves the negotiation of multiple considerations or factors, only some of which fall purely within the expertise of computer science. The discussion proceeds by first showing how this distinction arises across the lifecycle of AI system development and considering the role of senior leadership in directly or indirectly influencing these decisions.

This process of negotiating multiple considerations is perhaps the clearest at the earliest stages of system development. This is because, although AI systems are becoming increasingly capable, they remain tools whose value is entirely instrumental. In other words, the value of an AI system derives entirely from its ability to perform some task or function. If that task or function is of considerable value or importance, then an AI system that makes it easier to accomplish, or to accomplish it more quickly, or at larger scale, or more equitably, will also have
considerable value. As a result, decisions about which AI systems to develop are inherently decisions about which goals or objectives to prioritize, whether and how accomplishing these goals or objectives will produce net benefits, which stakeholders stand to be most advantaged or disadvantaged in this process, and whether attending risks that cannot be eliminated can be reduced to a level that is acceptable and how that threshold is to be determined.

The terms, “benefits,” “values,” “stakeholders,” and “risks” are plural because different tasks can generate diverse benefits that advance distinct values, for a range of stakeholders, with different tradeoffs and implications for risk. For example, relevant benefits might include deterring fraud by ensuring greater compliance with regulatory requirements such as “know your customer,” reducing the frequency or duration of network outages, increasing the rate of transactions that can be processed in given time and so on. These are benefits in the sense that they produce a certain sort so value. For example, banks have a set of social responsibilities they are expected to fulfill. This includes taking effective steps to ensure that their infrastructure is not used to support a range of anti-social activities, including fraud, money laundering, or to support terrorist activities.28 29 Similarly, the services they provide are often socially essential. As a result, access to those services, from banking to lending, has a profound impact on the economic opportunity and achievement of individuals as well as their social standing—the extent to which they are recognized as having equal status or worth. As a result, banks have a moral and a legal obligation not to discriminate in their practices30 and a moral obligation to promote equity in access to their services.31 32 At the same time, banks are profit-driven enterprises and some of the benefits outlined above might decrease revenue by increasing costs or reducing the number of transactions per unit time.

30 https://consumer.ftc.gov/articles/credit-discrimination
31 Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 (Dodd–Frank Act) lays out the goal of preserving and promoting Minority Depository Institutions (MDIs) and assigns specific responsibility to the Secretary of the Treasury and other regulatory agencies for assisting MDIs since “Typically, MDIs serve economically challenged communities traditionally underserved by the banking industry and other businesses. [...] MDIs are uniquely positioned to create positive change in these communities [...]” https://www.occ.treas.gov/publications-and-resources/publications/banker-education/files/2021-report-to-congress-minority-depository-instit.html p. 5.
Promoting these benefits can produce value for different stakeholders. Reducing fraud and restricting access to resources for nefarious activities produces a social benefit. To the extent that the steps necessary to achieve these goals require time or resources, they may appear to come at the expense of profit and so conflict with the interests of shareholders. But such a perspective overlooks the extent to which pro-social corporate activities enhance the reputation of the firm and establish the warrant for social trust in its operations. Nevertheless, technical advances that facilitate compliance with these requirements might further mitigate both the appearance and the reality of such conflicts. Likewise, advancing the goals of equity ensures that essential social services are available to individuals and groups who experience, or are at high risk of experiencing, social exclusion, poverty, and lack of access to essential services.

It is useful, therefore, to think of the process of deciding which tasks or functions an organization might use AI to accomplish as analogous to creating or selecting an “AI development portfolio.”33 Because technical expertise is limited and the set of tasks or functions that an organization might perform is potentially quite large, the challenge is to select tasks for this portfolio that will generate sufficient value, for the diversity of stakeholders whose interests are impacted by the banking sector, while ensuring that individual and cumulative risks can be mitigated or managed and remain within reasonable bounds.

The process of constructing such a portfolio involves identifying tasks that have two properties: it is feasible to accomplish them effectively, efficiently, and equitably with AI and they are likely to generate sufficient value for some set of stakeholders that they should be given priority over other available alternatives. At one extreme, the question of technical feasibility hinges critically on specialized expertise about the capabilities and limitations of current tools and methods in computer science. At the other extreme, decisions about how to prioritize the mix of values that might be generated for different stakeholders by alternative AI portfolios, and how these should be balanced by attending risks that might be generated for impacted stakeholders (recognizing that efforts to generate one type of benefit for one type of

33 Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard business review, 96(1), 108-116. P. 113 they discuss the value of creating a “prioritized portfolio of projects” with an emphasis on ensuring that projects have sufficient business value and can benefit from the application of AI systems. The main difference is that the present analysis emphasizes the wider range of considerations that determine the value of a task.
stakeholder might introduce risks that are born by different stakeholders), requires the interplay of social, ethical, and business considerations that fall outside the special expertise of computer science, broadly construed.

Constructing AI portfolios that safely and effectively advance legitimate social, ethical, and business goals requires an accurate conception of the capabilities of AI systems and the willingness and ability to use those systems only for tasks that advance these objectives. Inaccurate perceptions of the capabilities of AI systems can produce two types of failure. Overestimating these capabilities can expose stakeholders to risks associated with incorrect judgments such as false negatives or false positives. Underestimating these capabilities can result in opportunity costs associated with the continuation of practices or procedures that are less reliable, efficient, accurate, or equitable than an attainable alternative. Both failures result from mis-calibrated expectations grounded in the failure to communicate technical information the capabilities and limitations of AI systems to decision makers who may lack this technical background.34

In the third quarter of 2021 the online real estate company Zillow reported a $304 million inventory write-down after implementing an AI system to automate the process of purchasing homes with the goal of quickly updating the properties and reselling them for a profit. According to CNN, “Zillow CEO and cofounder Rich Barton explained the shuttering of Zillow Offers by citing ‘unpredictability in forecasting home prices’ that ‘far exceeds’ what the company had expected.”35 It is likely that part of this challenge involves anticipating how an AI system will interact with, and potentially alter, complex social dynamics. In particular, sellers have specialized knowledge about the quality of their home that will influence their decision to show it to buyers or take advantage of an offer from an intermediary. This creates the potential for an adverse selection effect for the intermediary.36 Part of the responsibility of senior leaders is to ensure that their business practices, including the deployment of AI systems, reflect an

accurate understanding of the social dynamics in which they take place. In this case, the costs of a mismatch between the business case and the capabilities of the developed system were born by the firm.

Even if all stakeholders have a perfect understanding of the capabilities and limits of all relevant AI systems, constructing AI portfolios that best advance legitimate social, ethical, and business goals requires decision makers to prioritize the relevant values, for the relevant stakeholders. Failures at this level can occur when decision makers fail to give sufficient weight or priority to one or more of these legitimate goals, fail to give sufficient weight or importance to generating specific types of value for relevant stakeholders, or fail to give sufficient weight or importance to the legitimate interests of affected stakeholders. In all these cases, the relevant considerations fall outside the technical expertise of computer scientists and involve considerations that fall into social, ethical, or business domains.

Non-Technical Aspects of AI Development Across the Lifecycle

Deciding which tasks to accomplish with AI is only the first step in the lifecycle of AI system development. Additional steps in this lifecycle include data acquisition and use, model training, model testing and validation, and creating the practices and procedures necessary to ensure a safe, effective, and equitable deployment. Although these stages rely heavily on expertise in AI and computer science, they also require the integration of social, ethical, or business considerations that fall outside the expertise of computer science. The point of the present section is merely to illustrate this interplay of different considerations. We turn to the question of who is responsible for these decisions in the following sections.

Data. Whether an AI system can be developed to perform a desired task depends, in part, on a range of issues related to the data needed to build such a system. These issues also vary in the extent to which they require specialized knowledge from computer science. For example, whether an organization can secure legitimate access to such data, or whether the data that an organization already possesses can be used for a particular purpose, hinges on ethical and legal issues involving privacy, confidentiality, and the specifics of how that data will be used and shared.
Similarly, whether available data is adequately representative of the larger population or populations likely to be impacted by the operation of an AI system requires a determination of what the relevant categories or features are that constitute a representative sample. But what categories or features of a population are relevant will depend, in part, on the duties, obligations, goals, and values of relevant stakeholders. Historical patterns of inclusion or exclusion are a social dynamic that affects the rights and interests of marginalized or excluded populations. These dynamics can produce datasets in which marginalized groups are underrepresented—there is less data about them—or in which data that is captured is biased. For example, biased lending practices can produce patterns of data that reflect the effects of discrimination rather than the ability of marginalized groups to successfully deploy resources. As a result, a dataset that is representative of the activities of a firm over a given period may not be representative of individuals from groups that have historically been excluded from the banking sector.

For example, in 2018 it was revealed that Amazon had shuttered a project designed to automate the screening of job applicants. The system was trained on previous resumes that had been submitted to the firm but was found to systematically deprioritize applications that had indicators that the applicant was female. In this case, the goal was to speed the process of vetting large volumes of job applications at scale. But the system operated in a way that would have unfairly discriminated against female applicants. In this case, it was not feasible to reconcile the desired use case, given the data that was available, with important ethical considerations of fairness and equal respect.

Models. If stakeholders regard the available data as fit for purpose, then a range of decisions must be made about the operating characteristics of the model that will be generated from that data. Although this is likely to be the most technically complex aspect of system development, it nevertheless involves decisions that fall outside the narrow expertise of computer science. For example, determining what constitutes an acceptable level of accuracy

and an acceptable rate of false positives and false negatives requires an assessment of how serious the consequences of these outcomes are likely to be for impacted populations. These decisions often require balancing or reconciling potentially competing goals. A system that sought to reduce fraudulent activity to the greatest extent possible would likely result in an increase in false positives, restricting legitimate financial activities of a range of stakeholders. Such outcomes might be particularly severe if they fall on individuals or groups that are already in a precarious financial situation or who already face significant hurdles to participating in financial markets.

Decisions about where to set thresholds on these features of a model implicate larger social responsibilities related to fair access to basic social services, equity in the redress of histories of social exclusion, how risks are distributed across social groups, and how the risks of any resulting model impact the overall risk-profile for a firm. Technical decisions about features of AI models involve, implicitly or explicitly, decisions about how to reconcile the broader objectives of a firm.

For example, the insurance firm Aflac has deployed an AI model to automate the approval process for simple insurance claims that don’t require proof-of-loss, as when a claim is submitted for a regular checkup or screening. Claims that fall into this category can be approved by the system without human oversight. However, the AI system cannot automatically process a denial—recommendations for denials must be evaluated and processed by a human. In this case, the firm assumes the risks from false positives which it hopes will be offset by the gains in efficiency from being able to process simple claims quickly, at scale. This sensitivity to the impacts of false positives stands in contrast to recent cases in which citizens in states such as Rhode Island, Idaho, and Michigan have alleged that automated systems for reviewing claims for federal or state benefits programs or for detecting fraud in such programs have wrongly denied them access to social services or wrongly penalized them after systems falsely accused them of fraud. In these cases, victims allege that systems automatically and incorrectly denied

41 https://spectrum.ieee.org/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold
42 https://time.com/5840609/algorithm-unemployment/
vast numbers of claims and that the benefits of false positives have accrued to cash-strapped
state governments while the burdens have fallen on individuals whose precarious social
circumstances entail that they stand to experience life-altering consequences from mistakes.

Testing. Once a model has been developed, it must be tested or evaluated to confirm
that it can perform the desired task within the specified parameters of accuracy, reliability, and
safety. Here again, although the work of designing and carrying out testing and validation is
highly technical, decisions regarding which methods to use and what standards to adopt require
judgments that go beyond the purview of technical experts and that impact the interests of
other stakeholders in ways that are likely to influence these decisions. Different approaches to
testing and validation require different amounts of time and resources to complete. As a result,
decisions about project timelines and funding can profoundly impact the approaches that are
used for model testing and validation. But these approaches have different degrees of reliability
and are susceptible to different kinds of bias. As a result, decisions about which methods to use
affect the profile of risk end-users face in implementing a system. To see this, it is helpful to
canvas four approaches to testing or validating an AI model and to provide some illustrative
examples of how these approaches can affect the residual risks that remain after testing.

For example, it is common practice to take a large dataset and divide it into two subsets,
one to train the AI model and another to validate the performance of the assembled model.
This practice of *internal validation* has the advantage of being inexpensive and expedient—
because the data used for validation are already on hand, models can be tested against the
holdout set quickly and efficiently. However, these virtues are often purchased at the expense
of external validity. The initial dataset used for model development and testing might fail to
capture features or dynamics of the real world in a way that will affect system performance in
practice. Because this is a problem that affects the initial dataset, it is present in both the
training set and the testing set. Performance measured on the latter may present a more
optimistic picture than how the model will perform in deployment. If, during the model
development stage it was determined that a particular threshold would need to be reached on
a given metric (for example, sensitivity and specificity) for a model to be worth using in practice,
it might be easier to demonstrate that a model can achieve such targets on a holdout dataset than to achieve these results in the real world.

One way to avert problems of this kind is to test the performance of a model on an independent dataset. This practice of *external validation* has the advantage of evaluating model performance on a dataset that was gathered by different parties or using different methods, perhaps at a different time or place. The more that these aspects of the independent data set differ from the training set, the more likely it will be that data capture regularities of the underlying data-generating process rather than arbitrary features of the way a dataset was composed. Even so, it can still be the case that the independent dataset and the training set contain correlated features that models use to achieve a level of performance that is unlikely to be replicated in deployment, where such features will be absent. For example, even if they are created by different groups at different times, each dataset might reflect a biased sample of cases because researchers in each group share a preference for “clean” examples, so that model performance is unlikely to generalize to real world cases where data is messier. Or the datasets might be assembled by groups in high-income settings with a similar infrastructure reflecting the behavior of more affluent parties so that the resulting datasets share features that are unlikely to generalize to lower-income settings with less robust infrastructure, involving the behavior of less affluent parties.

These problems might be averted if a system is tested, not on an existing dataset, but on data that is gathered from the setting in which the system is intended to be deployed. This can be done, for example, by running an AI system in “silent mode” where data from the deployment context is fed into the system in real time and its predictions are recorded and then compared to the predictions generated by the current approach to the problem, if there is one, and to the outcomes that eventually materialize, if these are known. Because this approach uses data from the real world in the context in which the system will be deployed, it can avoid biases introduced into datasets from the process of curation or from differences between the context of data collection and the context of deployment.

---

Silent trials work well in cases where AI systems are performing actuarial functions—where they make predictions about the world without intervening in the underlying data generating process. For example, if an AI system running in silent mode predicts that a network is likely to experience an outage soon, then the accuracy of that prediction can be evaluated by waiting to see whether an outage occurs.

Silent trials are less well-suited to cases where the goal is to intervene in a system to avert bad outcomes or promote better ones. If we do not want to predict network outages, but to avert them, then we need to know not simply whether a system predicts an outage, but we want to compare cases in which an AI system predicts an outage and we intervene, to the situation in which it predicts the same outage and we do not intervene. After all, if our goal is to avert outages, then a system that can predict them but not help us to avert them may be of little use. And it can be the case that a system predicts an outage, and we intervene, and no outage occurs, not because our intervention was effective, but because the prediction was faulty.

When what we ultimately care about is our ability to use an AI system to intervene in the world, the best way to measure the added value of using the AI system is often to conduct a randomized controlled trial (RCT) where the effects of implementing the AI system are compared to the effects of the current practice for achieving the same ends. RCTs are often considered the gold-standard for producing evidence of causal efficacy for an intervention. But they are also expensive and time consuming.

Ideally, models will be tested and validated using strategies that best align with the model use-case and the risks associated with its use. However, decisions that leaders make regarding development timelines and budgets for research and development can limit what is feasible, effectively representing a de facto decision regarding system reliability and model risk.

Once the results from this process of testing and validation are complete, decisions must be made about how best to proceed. In particular, this process can reveal weaknesses, deficiencies, risks, or failure modes for systems and one response can be to try to improve the system so as to eliminate, mitigate, or otherwise address these problems. In this case, decisions often must be made as to the source or origins of problematic aspects of a system performance.
and the best strategies for addressing them. Another response is to accept the performance of the system as developed and move it into deployment, perhaps with a notification or disclosure of its capabilities, risks and so on. A third option is to discontinue work on a system if it does not meet acceptable benchmarks for performance. Although the question of how a system’s performance compares to benchmarks for acceptability might be a technical question, the specification of those benchmarks and decisions about what to do considering these results involve considerations that range beyond narrow, technical expertise.

Deployment. Finally, the process of deploying an AI system requires the deliberate integration of an often complex computational tool within a larger workflow. Very often, any benefits that a system is designed or anticipated to produce depend crucially on the ability of stakeholders to implement it in actual practice within conditions that ensure accurate, reliable, and equitable performance. These conditions include having practices and procedures in place to ensure that the system is used to perform only the specific task for which its performance has been validated, to ensure that context and operating conditions remain similar to those under which system performance was validated, and that ensure that system outputs are incorporated into or propagated across the appropriate workflows necessary to realize beneficial outcomes and to reduce residual risks.

Ensuring that AI systems are incorporated into workflows in ways that ensure effective, efficient and equitable performance requires education, training, and in some cases the reconfiguration of human practices. Effectuating these conditions in practice requires buy-in from affected parties, including a willingness on the part of key stakeholders to trust the system in question so that they are willing to implement it and to act on its outputs, under appropriate conditions. Most of the key steps in this process involve decisions that range beyond the technical expertise of computer science.

Senior Leadership as Direct Stakeholders

The previous two sections outline some of the decisions that arise at key choice-points in the development of AI systems that involve considerations for stakeholders that fall outside of the narrow purview of technical experts in AI or computer science. The next point to note is
that senior leadership often are, or ought to be, among the stakeholders involved in making these decisions.

Senior leaders are often involved directly in decisions relating to the construction of a firm’s AI portfolio, whether this portfolio is constructed through internal system development or the acquisition or procurement of AI systems from external vendors. This is most often the case when decisions about whether or not to invest in AI constitute strategic decisions that are recognized as likely to impact core aspects of a firm’s operations. Such cases include investments in AI that: constitute a strategic change in the business model of a firm, require significant costs or allocation of capital or resources, are anticipated to produce significant benefits or returns for the firm, that are anticipated to require significant shifts in the workforce or organization within a firm, or that are anticipated to have significant consequences for the risk profile of the firm.

When the development or procurement of AI rises to the level of a strategic decision, then senior leaders are responsible for ensuring that the desired value proposition is feasible and worth the associated expense and risks. Such a decision cannot be made without assessing the considerations outlined above. This is most obviously the case with the relationship between the task the AI system will perform and whether successfully using AI for this purpose will translate into achieving the desired social or business goal with greater efficacy, efficiency, or equity. It is also easy to see how the viability of these goals hinges critically on whether data that is available or that can be acquired are suited for this purpose. In an economy that values and prizes data, it is not uncommon for senior leadership to make strategic business decisions with the goal of gaining or promoting access to data that can support data analytics or AI systems. Issues of model development, testing and validation may appear to be too narrow or technical to fall into the province of senior leadership but these issues are inextricably related to the viability of a system, the degree of risk associated with its use, and whether its use will advance the appropriate set of social, ethical, and business objectives.

In other cases, senior leadership ought to be involved in decisions relating to the construction of a firm’s AI portfolio because those leaders have special obligations or responsibilities that are impacted by these decisions. In banking and finance, for example,
Senior leadership bear ultimate responsibility for identifying, assessing, and ensuring that risks related to models the firm deploys remain within acceptable tolerances. In this role they rely on the expertise of parties who play various roles in the firm’s overall model risk management framework. Nevertheless, at the end of that process, senior leaders bear the ultimate responsibility for ensuring that model risk, both in terms of the risks profile for individual models and the aggregate risk from the joint deployment of all models used in the firm, have been appropriately managed and fall within acceptable tolerances.44

Senior Leadership and the Ecosystem of AI Development and Deployment

There will likely be cases where the development or use of an AI system does not rise to the level of a strategic decision. Nevertheless, the decisions of senior leaders can indirectly influence and shape the development or use of AI systems because senior leadership play a direct role in creating the norms, procedures, structures, incentives, and division of labor that constitutes the ecosystem within which AI development, procurement, and deployment take place. Senior leadership, therefore, has a responsibility to ensure that this ecosystem is consistent with and supports the ethical and responsible development or procurement, and use of AI.

Importantly, the strength of a firm’s commitment to ethical and responsible AI cannot be assessed solely from an inspection of its rules and regulations related to responsible and ethical AI. For example, it has been noted that the Enron Corporation had a substantial policy on business ethics. However, that document alone was not an accurate picture of the business ethics of the firm or of the considerations that drove actual decision making.45 The reason is that excellent policies on these issues can be overridden, subverted, or otherwise derailed by a larger norms, rules, practices, or procedures within the firm. As a result, the strength and credibility of an organization’s commitment to ethical and responsible AI must be assessed not solely as a function of the quality of local policies about the development or acquisition of

44 https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
45 https://www.fbi.gov/history/artifacts/enron-code-of-ethics
AI, but broadly in terms of how those local policies integrate into and are supported by, or are overridden and subverted by, the totality of practices within the organization.

Senior leadership thus have a unique and special role to play in ensuring the ethical and responsible development and use of AI. This role is unique because senior leadership represent the highest decision-making authority within an organization. They are responsible for decisions that have the most comprehensive scope within an organization, since decisions that are made within individual units can have a cumulative or synergistic impact. They bear responsibility for setting goals and priorities and for reconciling potentially conflicting values and objectives. They oversee the division of labor in a firm, including the delegation of rights, responsibilities, and obligations and they set corporate procedures that include incentives for advancing strategic goals, sanctions for breaching norms or constraints, and enforcement mechanisms for ensuring that the conduct of personnel aligns with both objectives and constraints.

Their role is special in relationship to AI because this technology often relies heavily on the distribution of knowledge and labor across different elements of an organization. As a result, the ability of an organization to satisfy the various requirements for ethical and responsible AI depends critically on organizational arrangements that coordinate the activities of relevant parties. This coordination involves (a) ensuring that all relevant responsibilities related to ethical and responsible AI are allocated to appropriate personnel, (b) ensuring that these responsibilities are tied to clear and, where possible, measurable criteria or benchmarks for success or failure, (c) ensuring that personnel are empowered to discharge these responsibilities, (d) in an environment in which relevant criteria for success or failure are used to evaluate the performance of individuals, groups, and systems, and (e) that processes are in place to identify, manage, and resolve conflicts.

Senior Leadership: Governance, Accountability, and Responsibility

AI system development takes place within a larger ecosystem that consists of actors who divide labor and distribute responsibility for the purpose of advancing a set of goals or objectives. Governance should be understood as the process of clearly stating relevant ethical,
legal, and business objectives and constraints, delegating responsibility for meeting these objectives within these constraints and creating a system of accountability to ensure that actors within this ecosystem adhere to sound practices, work to advance the relevant goals, and abide by the relevant constraints. In this context, a system of accountability refers to a system of norms, rules, roles, and incentives designed to ensure that every duty, role, or responsibility necessary for ethical and responsible development or use of AI has been allocated to a specific individual or group, that these individuals and groups face incentives that align with the furtherance of these ends within relevant constraints, and that mechanisms are in place to reward compliance and to sanction conduct that breaches a relevant norm or contravenes a relevant responsibility.

The responsibility for governance, accountability and responsibility falls directly to senior leadership. Senior leaders directly determine the system of governance within a firm and have the power and responsibility to shape the system of accountability and the delegation of responsibility. This responsibility plays out at two levels.

On what I will call the local level, leadership is responsible for setting up a system of governance and accountability to ensure that AI system development satisfies the requirements of ethical and responsible AI. Elements of this process may be delegated to managers, technical experts, advisors or consultants, but the ultimate responsibility for ensuring the soundness of this system is the responsibility of senior leadership. This set of policies and practices is “local” in the sense that they speak directly to aspects of AI system development and the operationalization of relevant norms and practices. The design and implementation of local systems for responsible and ethical AI has received the lion’s share of attention in the literature on ethical and responsible AI.

Perhaps paradoxically, attention to these local norms is insufficient to ensure that the objectives of this local system are advanced in actual practice. The reason is that local norms and practices are embedded within a larger ecosystem, governed by a larger division of labor and designed to advance potentially distinct goals and objectives. As a result of this embedding, sound norms, rules, policies and practices at a local level can be swamped, diverted, undermined, overridden, or otherwise subverted by broader norms within an
organization or by a faulty division of labor across the full process of system development and deployment.

On the “highest-level,” senior leadership is responsible for ensuring that the local governance structure directly relating to ethical and responsible AI development and deployment are imbedded within the larger governance structure of the organization in a way that preserves the relevant goals and constraints, ensures continuity in the division of responsibility, and continuity in the system of accountability for ensuring that each party to the division of labor and responsibility faces incentives to act in compliance with or promotion of the requirements of ethical and responsible AI.

Without understanding factors that are necessary to promote, or that can threaten, the ethical and responsible use of AI, it is unlikely that senior leadership can create, maintain, and continuously improve a system for the local governance for AI development and use and ensure that these requirements are preserved and advanced by the larger culture, norms, incentives, and practices of the firm. They can rely on the judgment and expertise of technical experts, unit managers, and other experts when it comes to specific aspects of the AI development or deployment process. But they bear ultimate responsibility for, and must be able to assess, the governance system that is implemented within the firm and so must understand how technical issues that fall into the province of area experts interact with the comprehensive set of relevant social, ethical, and business considerations.

At the extremes there are two archetypes for structuring the governance of AI in a firm: centralized and federated approaches. Centralized approaches delegate to specific leaders, such as a Chief AI Officer (CAIO), responsibility for ensuring that the firm’s AI strategy aligns with its business goals, that its development and procurement processes follow ethical and responsible AI practices, and that these practices carry over into deployment and use. One advantage of this approach is that there is a clear delegation of responsibility to specific parties whose role within the firm is to ensure ethical and responsible practices across the lifecycle of AI system development or procurement and deployment.

A key challenge for centralized approaches derives from a central message of the preceding analysis, namely, that the ethical and responsible development or procurement and
deployment of AI is a function of myriad decisions at numerous choice-points in the AI lifecycle, only some of which fall under the purview of technical experts. Similarly, the strength and credibility of a firm’s commitment to ethical and responsible AI is not solely a function of its policies and practices directly related to this set of concerns. Rather, it is a function of whether and how those local policies and practices are integrated into and supported by, or undermined and usurped, by the firm’s larger business culture. As a result, in large firms where numerous AI systems are developed and deployed across the workflow of the enterprise, the centralized approach risks either being ineffective or requiring vast aspects of the firm’s operations to be channeled through the CAIO.

In contrast, federated approaches to ethical and responsible AI seek to upskill and educate the workforce in a firm so that awareness of and responsiveness to factors that influence the efficacy, efficiency, and equity of AI systems is spread across the operations of the enterprise. One advantage of this approach is that it seeks to harness the local knowledge of the workforce to address technical and non-technical decisions that arise at the myriad choice-points in the lifecycle of AI development or procurement and deployment. Ideally, propagating awareness of, and commitment to, the requirements of ethical and responsible AI across the workforce of a firm will enable parties responsible for key decisions to surface relevant issues and identify appropriate responses as part of the normal business process.

Federated approaches face two key challenges. The first is to ensure that all relevant decision makers—including board members and C-suit executives—receive education and training sufficient to empower them to advance the goals of ethical and responsible AI. The second is to ensure that the division of labor, roles, norms, practices, and culture of the firm ensure that ethical and responsible practices are supported and maintained across the lifecycle of development and deployment.

Hybrid approaches that blend aspects of these more extreme archetypes are possible as well. Which approach is likely best for particular organizations will depend on several factors including the size of the organization, the complexity of its AI portfolio, and the experience and familiarity of personnel and leadership with the principles and practices of ethical and responsible AI. Critical to the success of any approach, however, is the commitment of senior
leaders to the thoughtful implementation of a coherent and comprehensive approach to
governance for ethical and responsible AI.

**Conclusion**

Discourse surrounding ethical and responsible AI often focuses on a limited subset of actors who influence the development and deployment of AI systems. Without an explicit and sustained focus on the responsibilities of senior leadership, it will be difficult to prevent or close loopholes in governance and to ensure that senior leaders are accountable for their role in creating an effective and equitable governance structure to ensure the ethical and responsible development and deployment of AI systems. The analysis presented here highlights the ways in which senior leadership directly and indirectly influence the ecosystem of AI development, procurement, and deployment, including their responsibility for evaluating the expected benefits and risks of the organization’s AI portfolio.