MIS Quarterly Executive

Volume 21 | Issue 2 Article 5

June 2022

Building an Artificial Intelligence Explanation Capability

Ida Someh

Barbara H. Wixom

Cynthia M. Beath

Angela Zutavern

MIS Quarterly Executive

Building an Artificial Intelligence Explanation Capability

Though companies are building artificial intelligence (AI) systems and integrating them into business operations, executives are concerned about AI's distinctive challenges (e.g., opacity) and seeking to develop new capabilities in response. We describe a new AIX explanation capability that companies must establish before their AI initiatives can thrive. This capability has four dimensions: decision tracing, bias remediation, boundary setting and value formulation. Together, these dimensions help organizations to address the challenges of model opacity, model drift, acting mindlessly and the unproven nature of AI.[1,2,3]

Ida Someh
The University of Queensland (Australia)

Barbara H. Wixom
MIT Sloan School of Management (U.S.)

Cynthia M. Beath
University of Texas at Austin (U.S.)

Angela Zutavern
AlixPartners (U.S.)

The Challenges of Explaining the Behavior of Black-Box AI Models

Artificial intelligence (AI) technologies are producing exponential improvements in the ability to find patterns in data, make predictions and recommend actions without explicit human instruction.[4] An emerging body of information systems literature dedicated to AI defines it as the "ability of machines to perform human-like cognitive tasks, including the automation of physical processes such as manipulating and moving objects, sensing, perceiving, problem solving, decision making and innovation."[5] The vast opportunities available from harnessing AI technologies are stimulating a modern-day gold rush for businesses and governments alike. For example, a recent forecast by International Data Corporation predicts that AI spending will

1 Hind Benbya is the accepting senior editor for this article.

2 The authors thank Hind Benbya and the members of the review team for their guidance throughout the review process. We also gratefully acknowledge research funding from MIT's Sloan Center for Information Systems Research (CISR) and support from CISR Research Patrons and Sponsors.

3 Results from the first phase of this research were presented at the HICSS practice track in January 2020. See Someh, I., Wixom, B. and Zutavern, G. "Overcoming Organizational Obstacles to Artificial Intelligence Project Adoption: Propositions for Research," in *Proceedings of the 53rd Hawaiian International Conference for System Sciences*, January 2020.

4 Sapp, C. *Laying the Foundation for Artificial Intelligence and Machine Learning*, Gartner, Inc., 2018.

5 This definition of AI is consistent with the *MIS Quarterly Executive* special issue on AI. See Benbya, H., Davenport, T. H. and Pachidi, S. "Special Issue Editorial: Artificial Intelligence in Organizations: Current State and Future Opportunities," *MIS Quarterly Executive* (19:4), December 2020, available at https://aisel.aisnet.org/misqe/vol19/iss4/4.

DOI: 10.17705/2msqe.00063

reach $97.9 billion in 2023, more than 2.5 times the spending level of 2019.[6]

Despite the promising AI trends and forecasts, pervasive AI adoption and use have proven problematic to achieve in practice, particularly for incumbent firms. Apart from a few examples from leading technology companies (e.g., Facebook's face-recognition system and Google's self-driving cars), most AI projects are experimental and have never been deployed in practice.[7] Recent academic and practitioner research links the lack of progress to the challenges of explaining the behavior of black-box AI models, such as those that apply deep-learning algorithms.[8] To determine outputs, the algorithms at the heart of such models rely on complex internal structures that are inscrutable to human decision makers.[9]

The opacity of the way in which AI models produce their outputs has been exacerbated by significant fallout and unintended side effects from recent examples of algorithmic-based decision making.[10] For instance, a system used in the U.S. for assessing prisoners' suitability for parole was shown to display bias against black people, and in the U.K., the algorithmic prediction of results in university entrance exams (which had been suspended because of the pandemic) left students from a poorer socioeconomic background at a disadvantage. In Australia, a government algorithmic debt-collection system inaccurately calculated and automatically deducted amounts from the most vulnerable citizens.[11] High-profile AI-driven mistakes such as these have left business leaders skeptical about the reliability of AI. How can they be sure that their organizations can use AI technology in acceptable ways and thereby reap the desired benefits for the various stakeholders?

Recent developments, such as DARPA's explainable AI[12] initiative, are beginning to tackle the issues arising from the opacity of AI-based decision making from a technical perspective. According to DARPA, these developments constitute efforts to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)." Additionally, behavioral research is being conducted to identify what constitutes an effective explanation and how individuals perceive and consume explanations.[13]

Description of Our Research

What remains less researched, however, is how organizations can manage inscrutable AI-based models that are being used in real-world situations. More specifically, the central problem for organizations is: Can they *manage* AI in a way that ensures the underlying models are properly understood, do not incorporate biases, cannot cause negative consequences and comply with emerging regulations (such as the EU's General Data Protection Regulation). Our research reported in this article investigated the distinct managerial challenges of deploying inscrutable AI in organizations and how organizations can nurture an ability to address the challenges. By establishing foundations in solid capabilities related to AI, organizations can more effectively deploy and sustain AI projects over time while also maximizing their more immediate AI-investment returns.

Through systematic engagement with the Data Research Advisory Board of MIT's Center for Information Systems Research (CISR),[14] we identified executives' concerns about deploying AI technology. Building on these concerns, we set out to understand how AI-related managerial

6 *Worldwide Spending on Artificial Intelligence Systems Will Be Nearly $98 Billion in 2023, According to New IDC Spending Guide*, Business Wire, Inc., September 4, 2019.

7 For more information, see Benbya, H., Davenport, T. H. and Pachidi, S. op. cit., December 2020.

8 Issues concerned with explaining the behavior of AI models are explored in: 1) Zhang, Z., Nandhakumar, J., Hummel, J. T. and Waardenburg L. "Addressing the Key Challenges of Developing Machine Learning AI Systems for Knowledge-Intensive Work," *MIS Quarterly Executive* (19:4), December 2020; and 2) Asatiani, A., Malo, P. Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. "Challenges of Explaining the Behavior of Black-Box AI Systems." *MIS Quarterly Executive* (19:4), December 2020.

9 ibid.

10 Mayer, A.-S., Strich, F. and Fiedler, M. "Unintended Consequences of Introducing AI Systems for Decision Making," *MIS Quarterly Executive* (19:4), December 2020.

11 Rinta-Kahila, T., Someh, I., Gillespie, N., Gregor, S. and Indulska, M. "Algorithmic Decision Making and System Destructiveness: A Case of Automatic Debt Recovery," *European Journal of Information Systems*, September 2021.

12 Turek, M. *Explainable Artificial Intelligence (XAI)*, Defense Advanced Research Projects Agency, available at https://www.darpa.mil/program/explainable-artificial-intelligence.

13 See, for example: 1) Miller, T. "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence (267), February 2019, pp. 1-38; and 2) Afrashteh, S., Davern, M. and Someh, I. A. "Enhancing Fairness in Algorithmic Decision-making through Perspective Taking," *ECIS 2020 Research-in-Progress Papers* (61), 2020.

14 For information on MIT CISR's Data Research Advisory Board, see https://cisr.mit.edu/content/data-board.

Table 1: AI Projects Have Four Distinct Challenges

AI Challenge	Description	New Requirement
Model Opacity	Lack of clarity and understanding about the mechanics of models	Ensuring the compliance of the models
Model Drifting	Because AI models learn from the data on which they are trained, they are highly susceptible to technical and functional bias	Building representative models
Mindless Actions	AI models make decisions and perform tasks without being able to understand their limitations or judge the consequence of their actions	Ensuring the models are applied safely
Unproven Nature of AI	When, how and for whom AI will create value is often quite unclear and short-lived, and therefore needs constant management	Sustaining value from the models

challenges were being addressed by exploring a selection of AI projects in various states of deployment. By drilling down into two particular cases involving large-scale AI-based solutions—at the Australian Taxation Office and General Electric—we identified the practices employed to tackle AI challenges and how these organizations had systematically acquired and accumulated foundational knowledge that contributed to higher-level capabilities. (The research process is described in more detail in the Appendix.)

The investigation of these practices led us to identify a new capability that companies are building to facilitate pervasive AI adoption and use. This capability, which we call the "AI explanation capability (AIX),"[15] assists project teams in developing models that are compliant, representative, reliable and value-generating. AIX allows organizations to build confidence that AI will do the right thing in the right way at the right time.

In this article, we describe the AI-based solutions deployed at the Australian Taxation Office and General Electric, and identify the practices that together built these organizations' AIX capabilities. Based on our analysis of these two cases, we provide recommendations for leaders looking to exploit AI for significant advantage over time.

Distinct Challenges of AI Bring New Requirements

Most executives view their organization's use of data as a journey that moves, over time, from reporting and dashboarding to descriptive analytics projects, and then to predictive activities. The evolution toward more advanced data-related initiatives results from organizational learning and capability development in response to emerging technologies. Nevertheless, business and IT executives were telling us that "AI projects feel different." They are harder and take longer. They pose new challenges and demand new kinds of expertise, resources and capabilities. Because of these new challenges, the AI journey has not been proceeding as fluidly and seamlessly as previous journeys involving data. To address this perplexing concern, we undertook online executive-level discussions[16] with members of the Data Research Advisory Board in the first quarter of 2019. These discussions revealed that AI projects have four distinct challenges that set AI apart as a unique data phenomenon. These challenges, and the new requirements they bring, are summarized in Table 1 and described below.

15 We opted to conceptualize this new capability as AIX rather than employ the notion of explainable AI (XAI), which is commonly used in computer-science literature, because AIX goes beyond technical transparency and also connects back to a long tradition of information systems literature on explanation. See, for example, Gregor, S. and Benbasat, I. "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," *MIS Quarterly* (23:4), December 1999, pp. 497-530.

16 A set of practice-based research propositions arising from these discussions has been published in Someh, I., Wixom, B. and Zutavern, A., op. cit., January 2020.

Model Opacity

> *"Fear of 'the black box.' We work in a very high-risk industry. It will be a long time before we leverage technologies that self-learn and limit or remove human interaction."* Member of MIT CISR's Data Research Advisory Board

AI draws on sophisticated computational mathematics and statistics that make it very hard even for some data scientists, never mind business people, to readily understand an AI model's mechanics or how inputs are turned into outputs. Advanced AI models such as deep learning lack traceability; that is, there is no way to "follow the dots" from start to finish, as the models back-propagate their learning across multiple large layers of a network. Despite their opacity, these models often perform better than many traceable alternatives. This means that organizations may have to make a trade-off between simple algorithms that are easily understood and more sophisticated ones that perform better but are inherently opaque.

The risk of model opacity is that an untraceable or hard-to-trace model could be functioning incorrectly or producing results that are wrong, or that it becomes impossible to explain to stakeholders why the model made a particular decision. Model opacity may therefore make it difficult to comply with the EU's General Data Protection Regulation, which requires that "meaningful information" about the mechanics of the model be communicated to citizens subject to decisions made by AI-based services. Moreover, using untraceable models is utterly prohibited in the financial sector. Thus, the requirement for AI project teams is to be certain that the AI model complies with relevant regulations and legislation. To do this, they must find ways to untangle the computations and mathematics at play and convey the "how" behind the results to those who need to consume and make use of the output.

Model Drift

> *"You have to figure out how to monitor the performances, the data drift, and so the model drift. With eight models, the human can do it, but with 2,400 models for one business initiative, the human cannot do it anymore."* Member of MIT CISR's Data Research Advisory Board

In conventional IT projects, domain experts who are specialists in the relevant field set the business rules, while the systems are designed to provide decision support. In contrast, AI algorithms learn from training datasets, and the resulting models contain the decision-making rules that create outcomes or results.[17] During the learning process, AI-based models can drift in new directions, following whatever data they are given—whether good or bad, suitable or unsuitable.

The risk in model drift is that results may be biased. The biases may be technical in nature (e.g., arising from unbalanced datasets) or functional (e.g., biased from datasets that teach the model to make discriminatory decisions). Even after deployment, there is an ongoing risk of biasing data being introduced inadvertently into the model's feedback cycle. Moreover, without periodic retraining, the results produced by the model will inevitably deteriorate over time as the real-world changes. The risk from model drift imposes a key requirement on AI project teams: they have to find ways to expose model decisions in pursuit of identifying and eliminating bias from the models' training. These efforts can help build models that are representative of their populations and subgroups.

Mindless Actions

> *"Our employees have a hesitation in trusting the result, direction or action the AI will take vs. the good old-fashioned brain power of a traditional marketer, sales rep, data analyst, ..."* Member of MIT CISR's Data Research Advisory Board

When AI is implemented and scaled up, algorithms handle certain tasks and decisions instead of doctors, engineers, marketers, financiers or other domain specialists. AI models, however, are only limited representations of reality, and by their very nature incomplete. There is always some level of error in attempts

17 Faraj, S., Pachidi, S. and Sayegh, K. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), March 2018, pp. 62-70.

to carry over learning from training scenarios to new real-world cases; this is sometimes called "the framing problem" (treating all the new cases as if they were identical to some case that was used to the train the model). Autonomous models are not aware of this problem. They are always acting mindlessly.[18]

The risk of delegating tasks and decisions to mindless AI agents is unexpected and unintended consequences, which could result in loss of money, respect, health or even life. Deploying autonomous algorithmic-based systems in the workplace has to be accompanied by sensible limitations on the use of the results produced by the AI model. The new requirement for project teams arising from the ability of AI systems to act mindlessly is therefore to ensure the model is applied safely. This requires the project team to have: 1) a realistic assessment of the decision situation's accuracy and confidence levels; and 2) a profound understanding of the domain's decision-making processes, standards, exceptions and risks. The team must also provide the means for translating the AI model's non-definitive output into meaningful messages that lead to correct actions by humans or automated processes.

Unproven Nature of AI

> *"'[Our executives] all hear about [AI], they want it, they think it is cool.' (direct quote from our Chief Customer Officer) But 'when push comes to shove, they are hesitant to take away investment from traditional forms of P/L spend and invest in AI."* Member of MIT CISR's Data Research Advisory Board

AI techniques have been maturing for decades, but only recently have synergistic advances in technology, data availability and data-science skills propelled AI into the mainstream. However, there is not yet a broad range of proven use cases, and significant risk remains in AI initiatives. Organizations still cannot truly evaluate the risk of financial losses, reputation damage, sanctions from regulators, etc., should their AI systems act in undesirable ways. There is also a lack of understanding on the other side of the equation: What business problems or opportunities can AI-based solutions address? To make matters worse, organizations are not sure how to assemble the right resources and capabilities for generating meaningful value from AI initiatives.

One of the main risks of dealing with an unproven or amorphous technology such as AI is that it discourages leaders from making the required investments. Moreover, sloppy or uneven IT efforts can also yield negative consequences. To help organizations make the right investments, project teams need to show that AI is worth investing in and explain how value from AI can be managed and sustained over time, for all stakeholders. AI teams' messages need to resonate with a wide array of both internal stakeholders (e.g., technologists, system users, managers, executive champions and funding committees) and external ones (regulators, customers, etc.). AI's unproven nature imposes a new requirement on project teams to ensure that value from AI models is sustained over time.

To confidently create value from AI investments, organizations must actively manage all four challenges described above and address the new requirements arising from them. Below, we describe how the two selected AI case organizations built AI explanation capabilities that enabled them to meet these new requirements.

The Two Case Organizations and their AI Projects

To illustrate how organizations are addressing and managing the challenges of AI in practice, we describe large-scale AI projects at the Australian Taxation Office (ATO) and General Electric (the key details of these cases are summarized in Table 2). Both of these cases involve the deployment of complete AI solutions by well-established and substantial organizations, and we have corporate approval to discuss them in public. Both projects also used an array of practices that not only tackled the challenges of model opacity, model drift, mindless actions and the unproven nature of AI, but also built AIX capability that underpinned safe, large-scale deployment of AI.

18 Salovaara, A., Lyytinen, K. and Penttinen, E. "High Reliability in Digital Organizing: Mindlessness, the Frame Problem, and Digital Operations," *MIS Quarterly* (43:2), June 2019, pp. 555-578.

Table 2: Details of the AI Projects at the Australian Taxation Office and General Electric

Case Organization	Australian Taxation Office	General Electric
Sector	Government	Manufacturing
Employees (FTEs in 2019)	20,000+	205,000
2019 Financials	$426 billion in net tax collections	$95 billion revenue
AI Deployed for …	Checking compliance	Checking compliance
Initiator of AI Project	Corporate Digital Technology unit	A business unit
AI Use Case	Evaluate taxpayers' business expense claims for compliance with tax rules during the online claim-submission process	Evaluate documents containing details of contractor practices for compliance with GE safety principles during the contractor onboarding process
Approach to Machine Learning	Unsupervised learning	Supervised learning
Role of AI	Generate a behavioral nudge	Perform document evaluation
Nature of Business Change	Real-time nudging functionality added to the existing online tax filing process	Add-on created for the pre-existing contractor-onboarding process
Bottom Line	During the 2018 tax year, nearly 240,000 taxpayers were nudged, resulting in expense-claim adjustments of approximately AU$113 million	By 2020, the add-on was supporting hundreds of GE environment, health and safety professionals involved in the onboarding process

After providing an overview of the ATO and GE AI initiatives, we describe the specific practices that the respective project teams employed to address the four challenges and achieve model compliance, representative models, safe model application and sustained value.

Using AI in the Submission of Work-Related Expense Claims at the Australian Taxation Office[19]

The Australian Taxation Office (ATO) is the revenue-collection agency of the government of Australia, responsible for administering the country's tax system and significant aspects of its superannuation system. The ATO launched a program in 2015, called Smarter Data, to increase its data analytics capabilities. This program consolidated the activities of more than 500 experts—data managers, data engineers, data analysts, data scientists and business analysts.

One pillar of Smarter Data's mission was to help reduce the tax gap, the difference between the amount the ATO collects and what would be collected if every taxpayer were fully compliant with the tax regulations. For personal income tax in 2015-2016, Australia's tax gap was estimated to be 6.4%, or AU$8.7 billion ($6.3 billion).[20]

One AI use case suggested at the agency was to reduce the tax gap from individuals' work-related expenses. Legitimate expenses, such as travel, clothing/laundry for work, self-education and purchases of work tools or equipment, may be offset against taxable income and thereby reduce the tax owed. More than eight million returns every year claim work-related expenses, accounting for nearly 48% of the estimated tax gap for individuals.

Recognizing that it would be impossible to manually check and verify eight million tax returns each year, the agency asked Senior

19 Material relating to the ATO case is based on Wixom, B. H., Someh, I. A. and Gregory, R. W. "The Australian Taxation Office: Creating Value with Advanced Analytics," *MIT Sloan CISR Working Papers* (447), Massachusetts Institute of Technology, November 2020.

20 Currency conversion rate as of September 2021.

Director of Data Science Ying Yang to boost the use of advanced analytics, with the aim of creating a better expense-claim process. Yang described the problem like this: "We are dealing with millions of people, which means we can't audit our way out. We don't have enough human resources to check everyone's claims." Yang decided to implement a solution that uses sophisticated integrated machine-learning techniques to perform ethical checking and verify the expense claims at scale.

Yang and her colleagues wanted to make the very complex process for submitting tax returns easier for citizens and, at the same time, help auditors perform more efficient and effective oversight: "[We wanted] to help our auditors prioritize their resource allocation so they can focus on those cases that really need their attention, instead of wasting time fishing in the big ocean for everything possible."

For this endeavor, Yang managed a team of data scientists who, while highly talented, lacked a deep understanding of taxation-related rules, laws and standards. To compensate for this lack of knowledge, the data scientists collaborated closely with the ATO's Individuals Market team, which has extensive knowledge of the highly nuanced domain of personal income tax.

The work-related-expenses project resulted in AI being incorporated into myTax, the ATO's online claim-submission system, for the 2017 tax year. The AI system generates a pop-up message that nudges the taxpayer in real time to check the numbers when a claim for expenses raises a flag. In the 2018 tax year, about 240,000 taxpayers (around 6.6% of myTax users) were nudged to review the "deductible" or "not-deductible" label for at least one specific work-related expense. This resulted in adjustments totaling AU$113 million.

Using AI to Assess Conformance of Contractor Documents with GE Health and Safety Standards[21]

GE's corporate Environment, Health, and Safety (EHS) team delivers company-wide governance and oversight in its area of expertise. In 2016, this unit identified a potential new opportunity: to proactively identify and prevent the possibility of high-risk contractor operations. In response, a team of EHS leaders formulated a set of "Life Saving Principles," or LSP standards, designed to guide work practices applied in high-risk operations.

With this starting point established, GE expanded its contractor-onboarding process to confirm that contractors had sufficient EHS programs in place, by evaluating each contractor's alignment with the LSP standards. Contractor onboarding was already an exacting and labor-intensive process, with hundreds of GE EHS professionals manually vetting the written policies of potential contractors (approximately 80,000 new ones annually). Adding LSP evaluations to this tedious process would bog it down even more. Furthermore, because many contractors were unfamiliar with GE's LSP standards, there was substantial back-and-forth discussion with them about how they would meet GE's criteria and requirements. In 2017, after the LSP standards were fully in place, Sandra Neale, an 18-year EHS veteran, wondered whether technology could more efficiently incorporate LSP oversight into contractor onboarding. She asked: "Can't a computer evaluate documents, assess risk using review criteria, and determine whether the criteria are met?"

A team from GE's Corporate Digital Technology unit was already focusing on digital transformation efforts for various corporate functions, including EHS. When Danny Slingerland, the head of this unit, learned of the business problem articulated by Neale, he tasked Vijay Ravi, a seasoned data scientist on his team, with assessing the viability of developing a machine-learning-based solution. Together, Ravi and Neale initiated an AI project that developed a contractor document assessment (CDA) tool, which was bolted on to the third-party cloud-based preliminary qualification application used in contractor onboarding.

By 2020, the CDA tool was supporting GE's EHS professionals. By pressing a button, the professionals initiate an LSP review of a document, and the tool analyzes the document and reports back as to whether the LSP criteria are likely satisfied or not. The EHS expert then decides whether to accept the CDA assessment,

21 Material relating to the GE case is based on Wixom. B. H., Someh, I. A. and Beath, C. M. "GE's Environment, Health, and Safety Team Creates Value Using Machine Learning," *MIT Sloan CISR Working Papers* (448), Massachusetts Institute of Technology, November 2020.

investigate the documents further or formulate a different assessment. If a contractor disagrees with CDA assessment, the document is reviewed by a second EHS professional.

The project team concluded that the AI-driven LSP review process was simpler and more consistent than any manual process. The CDA tool frees EHS professionals to focus their expertise on higher-value work, and also assesses contractors in a more standardized and streamlined manner.

How the Case Companies Addressed the AI Model Opacity Challenge

Managing Model Opacity at the ATO

For two key reasons, the ATO AI project team devoted significant time to addressing the model opacity challenge. First, as a government body, the ATO operates in a highly regulated environment, so its leaders had to communicate and justify its decision-making processes to regulatory bodies. Second, the ATO directly serves Australia's taxpayers, so its leaders had a strong sense of responsibility for creating taxpayer experiences that are fair and transparent.

Initially, the ATO project team chose a nearest-neighbor analysis technique that compared a work-related expense claimed with the amounts legitimately declared by taxpayers in similar circumstances (e.g., those in the same profession and having similar income) to identify claims that appeared to be excessive. This technique offered simple, readily understandable logic and traceable, justifiable results. Thus, the project team could explain the model's mechanics to internal and external stakeholders, educate them and engage with them. These explanations enabled stakeholders to provide feedback on the logic used and help to fine-tune it. Providing feedback reinforced stakeholder understanding of the AI decision-making process, and incorporating subsequent improvements proved important for the project's success, as described by Yang:

> *"One senior business director looked at the model and said, 'If you are closer to me than the remaining eight million people, it doesn't mean you are really close to me; you're just closer to me than the remaining. That made us go back and make adjustments. Through showing how the model works, a business director with no data science background understood and picked up on how we could improve the model."*

As the team integrated its work into the myTax system, it became evident that the nearest-neighbor technique did not support real-time execution. This technique was computationally slow, taking several seconds per tax return. This meant that the results could not be incorporated into myTax's online filing experience to guide taxpayers "on the fly." However, Yang and her team believed that identifying excessive claims after their submission would be costly both for tax officers and taxpayers. Instead, the project team developed a neural-network-based AI model to mimic the offline nearest-neighbor analysis results in a way that allowed real-time computation (taking only milliseconds per tax return).

The project team periodically ran the two models concurrently to make sure that their results were still properly correlated. Ross Barns, Director of Individuals Special Projects, characterized the process in this way:

> *"The training data for the neural network is the decisions from the nearest-neighbor analysis. Then we can compare the two decisions. We can show a curve demonstrating that the neural network is making similar predictions as if you are using the original nearest-neighbor analysis. That gave the business confidence that, yes, we are doing the same thing, just much faster."*

Thus, the project team applied a multi-model strategy that incorporated two AI techniques into its final solution. The nearest-neighbor technique provided transparent decision making, and the neural-network technique offered the necessary performance and scalability.

The multi-model solution not only proved practical, but was able to withstand significant oversight. Barns recalled, "We've had to explain [the solution] at the regulatory level, parliamentary level and court level. Because, obviously, someone will challenge a position,

[asking] 'why did I get selected?' We had to have a clear capability to explain intent and process." Regular reporting to senior management, regulators and other stakeholders reinforced the impression that the AI team was proactively creating positive benefits without neglecting negative external factors that might surface.

Managing Model Opacity at GE

GE's AI project team chose a natural-language-processing algorithm that could analyze text-based contractor documents. For the proof of concept, Ravi proposed a neural-network-based model that employed a deep-learning algorithm. The project team understood that the internal workings of the deep learning model would be incomprehensible to stakeholders but agreed that these limitations would be acceptable so long as team members had a firm and clear grasp of the model's behavior. The data scientists closely scrutinized recent academic research on deep learning, how the behavior of such models could be explained and corresponding architectures, to ensure that the trained model followed the latest standards. One approach that they employed to enhance the transparency of the model's predictions was to use text summarization techniques to present a combination of words that constituted the basis for the model's decision-making logic. The team invested in a graphical user interface that displayed excerpts from a document's text and communicated, with a selection of words, why a certain decision had been made.

However, because the GE application context was of an extremely sensitive nature and had "no-fail" risk tolerance, the AI team needed to do more to demonstrate that the model was making the right decisions. The team implemented an interface feature that supported manual tracing of the AI system's decisions for each contractor document. The CDA interface visually communicates whether a document has met or failed to meet a particular LSP requirement, with a blue or red indicator for passing or failing, respectively. Reviewers can drill down into the document under review to inspect it first-hand, to see the features and decision layers used for decision making and to see whether the model used the correct evidence.

How the Case Companies Addressed the AI Model Drift Challenge

Managing Model Drift at the ATO

The ATO invested time and energy to be sure that the AI model was representative of all Australians and that the datasets did not distort the algorithms during training or over time after deployment and use. This meant that the AI project team had to use data from historical audits selectively when building the model, because the historical data contained filing errors and a disproportionate number of non-compliance cases, which distorted the overall compliance rate. The team was also aware that the historical data might have had biases introduced by business rules, earlier goals and intentions, and auditors "auditing cases in one particular direction."

In addition to choosing training data selectively, the AI team opted not to label claims compliant or non-compliant at this stage because taxpayers might have different circumstances year-by-year that the model would not be able to capture. For this reason, the team chose an unsupervised learning approach to limit the possibility of introducing biases from human-origin labeling at the beginning of the learning process. The unsupervised learning algorithm was set up to explore the data freely and identify nearest neighbors for each taxpayer. If a taxpayer's claim was greater than the nearest neighbors' claims, then the taxpayer would receive a message to review particular aspects of the claim and decide whether to change the claim. However, the team invested heavily in checking the accuracy of the outputs (e.g., checking if nearest neighbors identified were in fact near).

The project team then devised a strategy for auditing the model's outputs systematically over time, with the aim of identifying any deviation from ground truth. Working with Smarter Data statisticians, the team created a stratified random sample of the higher-claims segment of the taxpayer population. This sampling approach enabled the team to identify a representative selection of cases. The sample was then subjected to manual auditing and the results were compared with the model's output.

By the 2015 tax year, the ATO AI team had tested the model and begun using it to provide auditors with pools of potentially high-risk cases. By checking these cases, the auditors aided in assessing and scrutinizing the model's results, which served as a feedback loop to the data scientists for improving the model. The auditors also gained experience with the model, which supported the gradual process of building confidence in its decisions.

The incremental approach was partly a result of the team's skepticism about enabling real-time learning. Rather than let the model learn in real time, from each citizen interaction, the ATO team trained the model on tax return data from the prior year and switched the real-time learning option off. The resulting system informed the tax-filing process for the next full tax year cycle. Because the model did not continuously learn and adapt, the process produced consistent results irrespective of when a claim was filed. Learning between years rather than "on the fly" guaranteed that the model's decisions treated citizens equally. Yang explained that "If our training data keeps changing, we would make different decisions for exactly the same case if it were run at different times. We always use last year's data so that our decision is consistent. That was a deliberate choice."

Managing Model Drift at GE

GE's AI project team used a natural-language-processing technique that required EHS experts to develop a list of words and phrases (called a "bag of words") that the algorithm would draw on for assessing LSP compliance. While Neale and her colleagues were tagging words in documents, Ravi ran the model over the growing set of training data. Early on, the team conducted a blind test to compare human agents' judgments of LSP compliance with the model's predictions. Somewhat surprisingly, one of these tests demonstrated that Neale herself, rather than an AI algorithm, was making errors in judgment. As she reflected:

> *"I had a hard time getting my head around that. One of the data scientists showed me the output, and I said, 'Well, no, that's not right.' And then when he came back with the parts of the document that were informing the model's decision, I realized that I had made a mistake. That process really convinced me of human error. I had developed the criteria. I had developed the bag of words. I've got a pretty good handle on evaluating the documents. So, if I'm making errors, the actual error rate of the manual process could be much higher than we expected."*

Given the uncertainty relating to decision making by both the AI system and humans, Ravi and Neale sought to understand what the model was learning and whether it was producing the right outcomes. To do this, they proposed formal side-by-side comparisons of assessments made by humans and by the AI system. Again, the graphical user interface played an important role. When indicating disagreement with the AI system's assessment, the reviewer could provide feedback on the output the model should have produced, which created more data points in the forms of tags and labels from which the algorithms could learn. All of the feedback generated was used later for retraining the model, and for educating evaluators as well.

The interface also provided access to the full review history, across all evaluators. It displayed a dashboard that monitored what was rejected, what was reviewed and the final decision. The cross-evaluator monitoring was also useful when Neale began to bring more document reviewers into the project. Once her core team had grown comfortable with reviewing contractor documents, she hired a third-party verification company to provide trained evaluators, with the goal of expediting the document-tagging process and monitoring model drift.

However, adding more people meant adding new human bias to the feedback data, which was made visible through the interface. Where two reviewers interpreted document text in slightly divergent ways, the cross-evaluator monitoring identified these as inconsistencies in assessments between humans, not just mismatches between the algorithm and human judgment. Armed with greater awareness of bias, the project team gradually reduced it by retraining both the model and the evaluators.

How the Case Companies Addressed the Mindless Actions Challenge

Managing Mindless Actions at the ATO

The ATO AI project team was keenly aware that machine-learning algorithms cannot fully capture the full complexity of a citizen's tax situation, and therefore could not be allowed to make a final decision on the legitimacy of any particular tax-expense claim. The team worked with the ATO's behavioral analytics team to integrate the trained model into the myTax system used by taxpayers to submit their tax returns online.[22] The behavior team encouraged the project team to use the AI model's predictions to nudge taxpayers toward making sure they were complying with the law. The vision was as follows: as taxpayers entered information into myTax, the model would assess claim discrepancies in real time, triggering a message prompting the taxpayer to check the entry when a discrepancy threshold was exceeded. The idea was that the nudge capability would limit the model's actions to only encouraging taxpayers to act responsibly. As Barns described it:

> *"The system is not designed to say what you've done is wrong. We absolutely make that clear in all of our documentation, that this is not a determination that what you put on your return is wrong. All we're saying is that your claim behavior is anomalous to the peer group against which we've benchmarked you."*

The project team felt comfortable that the nudging approach recommended to it was consistent with ATO principles, including that of "not policing citizens." The team also designed the final filing experience so that complex decisions were referred to humans. The agency's Deputy Commissioner Marek Rucinski expanded: "If something is going against a person, we wanted humans to intervene and make a final judgment as opposed to a machine making the judgment."

22 In the 2018-2019 financial year, about 30% of Australia's taxpayers submitted personal income-tax returns via myTax.

Managing Mindless Actions at GE

At GE, the fundamental objective for the LSP program was to improve safety, and so the project team was frequently asked if the model was safe enough to be deployed. The team was aware that a model's decisions would never be 100% accurate in practice and that they had to investigate when and where it would perform well. Neale set a requirement that the AI model should be extremely conservative in its evaluation of contractors' documents. The team identified situations in which suboptimal performance was acceptable and would not create a safety problem in the long run. Ravi explained:

> *"Any document that is classified as criterion satisfied when it is actually not satisfied is a real problem. If we allow the contractor to conduct that service, it could eventually become a safety and compliance issue. In the reverse situation, classifying a document as* ***not*** *satisfied that actually is satisfied, we may lose a contractor or irritate them, but that outcome does not cause serious harm or risk."*

In response to this requirement, the AI team put fail-safe controls in place to minimize false-positive results. These controls included making sure that documents were classified conservatively, performing manual validation if there was any doubt and continuously auditing the results produced by the AI model. The AI project team also addressed safety concerns related to the model's use by building it incrementally over time. Initially, the team focused purely on perfecting document assessments for a single LSP requirement. That phase involved manual inspection of every single document the model assessed. Only after the team was comfortable with the model's accuracy rates for that LSP did it expand the model to other LSP requirements. Gradually, the amount of manual auditing declined to about 5% of all documents.

The project team believed that it needed to communicate that the model results were imperfect to encourage human judgment. Collaborating with developers, the team built graphical and other interface elements that made the model-informed decisions and their

associated confidence levels as transparent as possible for the reviewers. One effective technique was to display the probability factor associated with each decision by the model: for each criterion, the display showed the probability that the associated assessment was accurate, which the reviewer could then use to determine if manual validation would be necessary. The dashboard also distinguished low-risk LSP review criteria from high-risk ones, in line with the project team's focus on eliminating inaccuracies associated with high-risk elements. In the production system, the reviewer can choose between manually assessing LSP compliance and using the CDA tool. Thanks to this control over the model's actions, the AI team felt comfortable bringing the solution into production and use.

How the Case Companies Addressed the Unproven Nature of the AI Challenge

Managing Unprovenness at the ATO

The ATO AI team was responsible for building a solution that generated value for the Australian government and its citizens. For the government, the AI team created an application that would help close the tax gap related to work-related expenses while reducing the cost of monitoring the compliance of expense filings. Historically, an auditor performed a tax assessment only after a taxpayer lodged a claim and was issued a refund. This process—known as a post-issue compliance check—was costly for both taxpayers and the ATO. With the new real-time nudging on work-related expenses, the ATO could intervene at scale because the AI system was reviewing the claims of all taxpayers who prepared their own return and filed it through the myTax system. This helped shrink the pool of cases selected for manual review by auditors; now, they only review high-risk cases.

There are also advantages for citizens. The AI team created a fast and appealing filing experience for myTax users that decreased process complexity without generating a feeling of "being policed." And, because guiding citizens and improving their experience was a central focus for ATO, the new filing experience gently prompted citizens who had simply made a mistake or were confused by the filing process to make adjustments, which, in most cases, also avoided the need for post-issue compliance checks.

For a more accurate assessment of the value of the real-time nudging in myTax, the ATO established a measurement team of behavioral analytics experts tasked with evaluating the outcomes of using the AI system. These experts created a method that captured the amounts of a return's expense claims when the taxpayer entered and exited the process. Once the tax return had been submitted, the difference between the two values was calculated to measure whether any claim changes could be attributed to nudging. Results from a study that included a randomized control group confirmed that the effect detected was indeed a result of nudging, proving AI's value to the senior leadership team.

Cultivating its new expertise in articulating, formulating and measuring value from AI was an important contributor to the ATO team's confidence in its solution. It made the team comfortable with communicating more widely about the AI efforts (e.g., in interactions with citizens, the regulator and other stakeholders).

Managing Unprovenness at GE

While Ravi was willing to consider machine learning as one possible way to support the EHS contractor-review process, he wanted to confirm the feasibility of this with a proof of concept before embarking on a full-fledged project. As he put it:

> *"I didn't want to commit and say, 'Yes, it is possible' and then three months down the line discover that we couldn't achieve the same, or higher, accuracy as manual reviews. Five years ago, efficient assessment through algorithms would not have been possible. Today, the availability of highly sophisticated algorithms and architectures do make it possible; however, these capabilities first require testing on a specific use case."*

Neale agreed to test the idea and mobilized a team to investigate whether a machine-learning solution was feasible. Adopting a minimum viable product approach, and focusing on a single LSP

Table 3: Summary of AIX Dimensions

Dimension	Description	Supports
Decision Tracing	The firm's ability to establish an understandable link between an AI model's inputs and outputs	Model compliance
Bias Remediation	The firm's ability to reveal and, thereby, rectify AI biases, unfairness, errors, flaws and other problematic discrepancies	Model representativeness
Boundary Setting	The firm's ability to explain an AI model's boundaries and the necessary scoping, limiting and interpretation constraints for actions, including consideration of assumptions, conditions, contexts and risks in using the outputs	Safe model application
Value Formulation	The firm's ability to explain how outcomes from an AI model influence decisions, processes and actions, coupled with how the associated changes will lead to a combination of cost, risk and value that produces value for the organization	Value from models

item, gave the team the ability to pivot quickly and mitigate the financial risk borne by GE should the machine-learning approach fail to work. Once the team leaders saw that each incremental development had yielded the desired outcomes, they were comfortable seeking continued investment for the following stages. The project team further validated the CDA tool before fully deploying it company-wide for all EHS reviewers by conducting a field test on GE's already-approved contractors to increase confidence in the tool's value.

As the project progressed, Neale regularly explained its purpose and potential benefits to EHS professionals who needed to understand the vision better. Neale believed that the most influential technique in her arsenal was to offer real-world examples of negative outcomes GE had experienced, coupled with explanations of how the CDA tool would have prevented them. She found that presenting impact metrics alone did not change hearts and minds, especially when funding was an issue, but storytelling did.

The AI project team established a permanent unit of eight people to monitor the technology-based components of the CDA process, conduct audits of oversight decisions, enhance the CDA tool and manage the machine-learning model as data, requirements and activities changed. The team continued to seek ways to make the CDA tool more effective. With the establishment of this ongoing oversight and management, EHS managers and GE's senior leaders alike felt comfortable that the AI solution would exert a positive influence on the company's standard contractor-onboarding processes.

The Four Dimensions of an AI Explanation Capability

As the ATO and GE projects illustrate, successful AI project teams rely on an array of explanation practices that support their efforts to address the four AI-related challenges. These practices indicate that AIX capability is something new and that organizational leaders must seek to develop this capability, via a focus on explanations. AIX capability is multi-dimensional, and includes making the inner workings of models understandable, creating explanatory interfaces to expose and enable rectification of biases in AI model outputs, setting boundaries for the safe application of AI models, and articulating and formulating a model's value for different stakeholders. Each of the four dimensions of AIX capability supports a specific goal and helps meet the AI-imposed requirements identified by executives, as summarized in Table 3.

Table 4: Summary of Decision Tracing Practices Used by the Case Organizations

Practices Used by the ATO
- Selection, correlation and concurrent use of multiple black- and white-box models
- Education of stakeholders in model logic
- Domain experts' involvement in adapting model logic

Practices Used by GE
- Communicating decision logic using text summarization techniques
- Domain experts' involvement in adapting model logic
- Tracing by users (aided by a graphical interface that allows users to scrutinize the evidence behind model decisions)
- Engaging with advanced AI research

The Decision Tracing Dimension of AIX Capability

Both the ATO and GE adopted practices that addressed the model opacity AI challenge. As such, these practices contributed to building the first dimension—decision tracing—of AIX capability, which we define as: *the firm's ability to establish an understandable link from the input taken by the AI model to that model's outputs*.

At GE, traceability was crucial because of the sensitive compliance context, which involved human and machinery safety. Domain experts were directly and deeply involved in preparing the training data and the model. These experts used a custom interface to trace every decision back to features and decision layers of the underlying document until the model routinely yielded no false positives. This user interface, originally built to communicate key information about decision logic to the experts, evolved into an important tool for scrutinizing the evidence behind the output.

To achieve the decision traceability required, the ATO (which operates in a highly regulated environment typical of a government entity) built trackable machine-learning models that were run concurrently to validate the high correlation between black- and white-box counterparts. The project team then complemented this technical approach with a humanistic/social approach that involved educating stakeholders about the AI model's mechanics, which meant they learned about the AI technologies involved, became savvier about the model's logic and provided useful feedback to the project team.

Both organizations continue to trace decisions: the ATO management team routinely examines the correlation between the nearest-neighbor and the neural-network models' results, and the human tracers at GE continuously review the evidence behind cases being evaluated.

The decision tracing dimension of AIX capability requires skills in selecting models, in educating stakeholders on the logic applied, in involving domain experts in adapting model logic and in designing tracing routines for users. Sound decision tracing practices ensure compliant models, and the managers interviewed pointed to the decision tracing aspect of AIX as key to exploiting AI in a wide range of use cases. The decision tracing practices used by the ATO and GE are summarized in Table 4

Bias Remediation Dimension of AIX Capability

The ATO and GE both dealt with AI model drift by continuously comparing human- and machine-origin decisions to identify bias and eliminate it. The practices used to address the model drift AI challenge built the second dimension—bias remediation—of AIX capability, which we define as: *the organization's ability to expose and redress AI models' biases, unfairness, errors, flaws and other problematic discrepancies*. The bias remediation practices included auditing of AI output, representative sampling and identifying bias by displaying a side-by-side comparison of AI decisions and human decisions.

The ATO's practices of stratified random sampling, which chose a representative collection of cases from the Australian taxpayer population for auditing, and combined unsupervised learning with auditor oversight of the results, made sure unbiased decisions were produced for

Table 5: Summary of Bias Remediation Practices Used by the Case Organizations

Bias Remediation Practices Used by the ATO
• Examining historical data for bias potential
• Human inclusion/exclusion strategies to reduce bias
• Auditing the AI model's output
• Representative sampling to detect model bias for different population groups
• Disabling the AI model's online learning
Bias Remediation Practices Used by GE
• A graphical user interface for articulating the models' outputs
• Side-by-side comparisons (for detection of both human and machine biases)
• Outsourcing of model drift assessment at scale to a third party
• Interactive feedback mechanisms and processes (enabled by the graphical interface) for ongoing evolution and validation

Table 6: Summary of Boundary Setting Practices Used by the Case Companies

Boundary-Setting Practices Used by the ATO
• AI-based nudging to limit the possibility of mindless algorithmic actions
• Strategy for referring complex cases to humans
Boundary-Setting Practices Used by GE
• Optimizing model performance to fail-safe
• A graphical user interface for communicating the AI model's confidence levels
• Strategies for keeping humans in the loop and for leaving room for interpretation
• Incremental and conservative development

different segments of the taxpayer community. These practices also led to new auditing techniques and a greater understanding of how to effectively build and use unbiased training datasets. Another important practice was to adopt an annual learning cycle for the AI model using the previous year's data to prevent the model from drifting as the current year's data came in.

At GE, a key element of bias remediation was the use of third-party assessors to audit the AI-based decisions, with comparison not just of human and AI decisions but also of the consistency of human decisions. By uncovering both machine and human errors, this practice enabled the model to be adjusted to eliminate bias (e.g., by changing its features). It also highlighted the need for new training to eliminate bias. The combination of bias remediation practices used at GE enabled the AI project team to develop a sophisticated understanding of user interface design for AI model management.

The bias remediation practices used by the ATO and GE are summarized in Table 5. These practices contributed to both companies being able to build representative AI models.

Boundary-Setting Dimension of AIX Capability

Both the ATO and GE dealt with the mindless actions challenge of AI systems by establishing very clear limits for the usage of the models' outputs, to prevent the results from leading to undesired impacts. These practices built the third dimension—boundary setting—of AIX capability, which we define as: *the ability to explain the AI model's boundaries and how AI-based actions can be scoped, limited or subjected to interpretation when the modes are integrated into workflows and get acted upon.*

The ATO used nudging instead of automated actions to influence expense-claim compliance, and GE's AI model provided guidance for human appraisal of contractors' documents by explicitly communicating confidence levels

Table 7: Summary of Value Formulation Practices Used by the Case Companies

Value Formulation Practices Used by the ATO
• Multi-stakeholder value proposition and management
• AI-assisted service design strategies
• AI value attribution and measurement
Value Formulation Practices Used by GE
• Continuous iteration and improvement
• Storytelling
• Marketing and communications strategies
• Business case formulation
• Professional development

(as probabilities) of its assessments. At both companies, humans could be called upon as needed for oversight, decision reversal and escalation/appeal of AI-generated decisions.

The boundary-setting practices used by the ATO and GE are summarized in Table 6. These practices enabled both companies to safely deploy their AI applications.

Value Formulation Dimension of AIX Capability

Both the ATO and GE used practices for testing, articulating, measuring, monitoring and communicating the value of their AI-based solutions to stakeholders, particularly project sponsors, investors and overseers. These practices built the fourth dimension—value formulation—of AIX capability, which we define as: *the firm's ability to explain how the given AI model's outcomes influence decisions, processes and actions, coupled with how the associated changes will lead to a combination of cost, risk and value that produces value for the organization.*

Incremental development and deployment, and credible impact measurement, were key in the development and evolution of a valuable AI solution for both companies. At the ATO, the AI project team engaged behavioral analytics and measurement units to shape and realize a desirable and sustainable value proposition. The GE AI project team learned in particular from increasingly extensive field-based application of the solution. The GE team also reached out to operations teams (such as developers and the owners of the contractor-onboarding process) for assistance in operationalizing the AI solution in such a way that it could sustainably deliver value over time.

The value formulation practices used by the ATO and GE are summarized in Table 7. These practices enable both organizations to continue to reap value from their AI systems and overcome the challenges related to the unproven nature of AI.

Building a Comprehensive AIX Capability

Our capability-focused approach to explaining the behavior of AI applications implies that dealing with the new challenges posed by AI demands new skills and knowledge that can be acquired either through training, systematic learning-by-doing, external sourcing or institutionalizing practices in policies. Our analysis of the ATO and GE cases provides insights into the nature of the AIX capability, formalizes its dimensions and makes available a set of example practices that project teams can draw upon, fruitfully use and extend. Some of these practices are synergistic; for example, user interfaces that can support multiple explanation dimensions. The practices used by the ATO and GE are evidence that not only are these organizations building an AIX capability, but that they will continue to learn by doing and thus accumulate additional AIX capability.

The AI explanation practices described above should be seen as building blocks for a higher-level AI explanation capability that can be shared throughout the enterprise. While companies are increasingly initiating AI projects, both our research and professional experience show that

many AI projects operate as isolated efforts, creating their playbook as they go. Organizations that adopt this approach fail to appreciate the "building-block potential" of their projects. Explanation practices are chosen to address specific project needs as they arise, with little coordination or learning across projects. As a consequence, new AI projects, often in other parts of the organization, tend to "reinvent the wheel," selecting or crafting practices from scratch.

However, our research suggests companies whose AI projects have reached full deployment and maturity are better positioned to build a comprehensive AIX capability. Leaders of such companies are beginning to view AI projects not as isolated efforts but as opportunities that collectively, over time, will help the company overcome AI-related obstacles and successfully deploy AI applications. They increasingly recognize that practices related to AI—especially explanation practices—can be repeatedly applied, carried over across initiatives or at least adapted to the needs of various kinds of AI projects. Hence, the managers at the helm of these companies proactively and regularly assess what the company knows about AI, and they use their AI projects for honing practices that can be accumulated and reused and are repeatable and adaptive. In effect, they are using AI project practices to build their AI explanation capability.

Recommendations For Building AI Explanation Capability

In this article, we have examined the unique challenges posed by AI (model opacity, model drift, mindless actions, the unproven nature of AI) and have identified the explanation practices required to address these challenges. Together, these practices build an organization's AIX capability. As organizational leaders gain an understanding of what AI explanations are necessary, and why and to whom these explanations must be made, we believe they can leverage the AIX capability to help change hearts and minds across the organization regarding the deployment of AI. Drawing on our cases, we provide four recommendations for organizational leaders for cultivating their AI explanation capability. Though these recommendations have been framed by the experiences of organizations with mature AI implementations, they represent general (yet powerful) insights that should assist with AI endeavors at any point along the journey—irrespective of the organization's starting state of AI maturity or experience.

1. Include and Engage with the Entire Organization to Build AIX Capability

Our research confirmed that AI project teams do a lot of explaining to an array of organizational actors and for many purposes. Domain experts explain existing business problems or opportunities to senior management to secure funding and other resources. Data scientists explain model mechanics to domain experts to establish trust in AI's viability; they also explain model mechanics to regulators to validate that a technique is permissible. Domain experts explain abnormal model outputs to domain managers to remediate possible domain knowledge gaps. AI project team leaders explain calculated impacts to senior management to justify continued funding and resources; they also explain measurable impacts to end users to incentivize users to buy into and adopt AI solutions. And the list goes on. Moreover, the explaining happens in different ways, for example through classroom education, PowerPoint presentations, proofs of concept, experiments, prototypes, assigned roles, direct involvement, tools, interfaces, reports, casual conversations and impromptu updates. Project teams should expect to explain AI to many different constituencies in a host of ways, with the overall goals of educating the organizational workforce, training them with respect to AI, and increasing their ability to be included and to meaningfully participate in AI-based model development, refinement and operation. Organizations should therefore ensure that their AI teams have the necessary explanation capabilities so they can explain to lots of people, in lots of ways and for lots of reasons, especially for boosting workforce education in AI.

2. Look Beyond the AI team to Assemble the Required AI Explanation Expertise

Many managers focus on staffing AI project teams with data scientists who can help translate domain expert needs into an AI-based solution. Data scientists, however, represent just one

area of expertise that AI projects require. Other competencies that may be less obvious but incredibly helpful for building AIX capability include user experience design, financial measurement and value formulation expertise, providing human-like explanations relevant to users, and the ability to detect bias and build representative data assets. Our interviews indicated that as AI projects evolve into solutions that need to be scaled and institutionalized, AI teams increasingly draw on expertise from across the organization to learn what to explain and how to communicate and consume it. In this way, they credibly and sustainably measure and communicate model impact, articulate how a model can mesh with established processes and systems, and guarantee that the model continues to receive data that reflects its dynamic reality.

3. Document Current Practices for Decision Tracing, Bias Remediation, Boundary Setting and Value Formulation

As AIX capability is new for many organizations, there is an evident need to take stock of best practices, some of which will be known because of past experience while others will need to be developed through organizational learning and development. Creating an inventory of current practices for decision tracing, bias remediation, boundary setting and value formulation provides a starting point for tuning and spreading best practices and building AIX capability. The inventory will also enable the organization to assess gaps and identify AIX development needs. Sharing and promoting good practices or institutionalizing them as policies and templates can promote AIX capability development. Moreover, reflecting on the continuing development of selected practices will promote the awareness-rich evolution of AIX capability. However, organizations also need to retrospectively examine each AI project and reflect on how it might improve AI-related explanation practices. Reflecting on the past can crystallize participants' experience as lessons learned for the future. Useful questions to ask are: What practices did the AI project reuse? What needed to be adapted? What new practices had to be created? Which of the practices employed could be drawn on and applied by other teams in future work?

4. Consolidate AIX Capability Development in a Central Group that Can Accumulate Expertise Over Time and Disseminate the Fruits of that Expertise at Scale

Both the ATO and GE had organizational structures that provided coordination and connection for AI projects. In essence, these structures helped both organizations to build their AIX capabilities. Three types of structures are commonly used to advance data-related organizational learning: an enterprise center of excellence, an enterprise-services unit and a community of practice.[23] At the ATO, the Smarter Data program served as an enterprise services unit for the provision of resources, while GE drew on a center of excellence structure in the form of the Corporate Digital Technology unit, which offered advisory support for corporate functions. We recommend that leaders select one or a combination of these scale-enabling structures and create a centralized unit for identifying helpful lessons, techniques and practices for AI explanation, and making them available across the organization to those working with AI. Centralized coordination will guarantee that the organization systematically captures lessons from the otherwise often localized efforts to build its AIX capability and disseminates the relevant lessons to inform organization-wide practice.

Concluding Comments

Businesses have good reason to be excited about AI opportunities: AI models execute certain tasks better than humans, preexisting processes and conventional technologies. AI enables the ATO to carry out population-level oversight, and GE uses AI for nuanced and sensitive, yet standardized and accountable, reviews of tens of thousands of contractors globally, without imposing unacceptable cost or risk burdens. The exploitation of the full power of AI in such companies is grounded in their creation of a specific AI-related capability—the ability to explain how AI models make their decisions. This

23 For more information on these organizational structures, see Someh, I. A. and Wixom, B. H. "Data-Driven Transformation at Microsoft," *MIT CISR Research Briefings* (XVII:8), August 2017.

AIX capability not only underpins success at the AI-project level but also provides organizations with the opportunity to become a successful player over time in the emergent algorithmic economy.

The unique needs of AI give rise to a multitude of explanation practices. Investigating explanation in terms of these needs, as we have begun to do in this article, will help leaders to better comprehend what needs to be explained about AI, why, and to whom. We believe that a greater awareness of AI explanation requirements will enable organizations to take full advantage of the AIX capability—in ways that help change systems and structures for the better and move both hearts and minds throughout the organization with regard to the deployment of AI.

Appendix: Research Process

The research for this article had three phases and was designed to investigate barriers to AI adoption. Phase 1 involved asynchronous discussions with executive members of CISR's Data Research Advisory Board about AI-related challenges and retrospective reviews of AI projects. Phase 2 comprised 100-plus structured interviews with AI professionals. In Phase 3, which is ongoing, we are conducting case studies of several organizations. The details of each phase are described below.

Phase 1: Identifying the Distinct Challenges of AI Systems

In Q1 and Q2 of 2019, two researchers initiated online discussions with the MIT Center for Information Systems Research's Data Research Advisory Board, whose 95 members are data executives in 67 large companies, headquartered around the globe. Each executive was asked to answer the following question on an online discussion board (with a period of one month to submit responses): *What are the top three impediments to AI adoption/consumption in your company?*

In all, 53 data executives from 50 organizations answered the question, a 75% response rate from the organizations represented (several non-respondents did reply to our request explaining that their lack of an answer was due to there being no AI activity in their organizations).

To complement the executive responses, the fourth non-academic author contributed a set of 85 AI project retrospectives that her firm had created with a broad set of client companies for knowledge development purposes. From this material, with all identifying details removed, one of the academic authors created a representative sample of six projects for the full research team to analyze. This purposeful sample enabled the team to identify how a wide spectrum of obstacles connected with AI had been overcome in practice.

Phase 1 of the research revealed the current state of AI and highlighted gaps in practitioner understanding. In particular, the difficulty of explaining how AI models reach their decisions consistently emerged as the main roadblock to AI's adoption. As a consequence, we decided to investigate and analyze this issue at a project level. Our aim was to develop a practice-grounded understanding of how companies globally are overcoming the challenges of explaining how AI models work.

To verify our findings, we conducted two webinars with the Data Board executives who participated in the first phase of data collection. This feedback was important for verifying the four dimensions of AIX capability.

Phase 2: Addressing AI Challenges with AIX Capability

From Q3 2019 to Q2 2020, the academic researchers collected data on 52 AI projects. The goal of this phase of the research was to investigate how AI project teams are addressing the new challenges posed by AI. The first two authors conducted 100 semi-structured interviews with 38 domain experts, 49 data scientists and 13 consultants at 48 companies. For 42 of the projects, these two researchers used paired interviews (i.e., a data scientist and a domain expert from the same AI project team were present during the interview). The interviews were documented via video recordings and later transcription. The interviewers followed two distinct protocols, one for the data scientists and the other for the domain experts. The questions for the former dealt with the development of AI models, deployment of AI solutions and scaling issues, and covered such concepts as explanations and trust-building. The interviews with domain experts focused on

how they engaged with and perceived the output from AI models and explanations of how the models worked, how they contributed to model development, and whether and how AI had changed work practices and skills.

The full dataset, in all its breadth, informed our development of the AIX capability concept. However, this article focuses, in particular, on further engagement with ATO and GE that we undertook to increase the depth of our case material, as described below.

Phase 3: Case Development

In Phase 3 of the research, we are engaging with a selection of companies in our AI project pool to develop detailed case studies. To date, we have written up four cases, including the ATO and GE ones, and are expanding this list. In developing the ATO and GE cases, we conducted four interviews with informants from both organizations. Though this might seem like a small number of interviews, they were embedded in a much broader process of engaged research, set in the context of a long-term relationship with each company where:

- We had access to company-specific material, such as archival documents and extensive public information
- We carried out multiple interviews for iterative clarification of our understanding, including several follow-up interviews for clarification of certain points
- The quotes included in the cases (and this article) were reviewed by the individuals concerned, and the case drafts were reviewed by the project teams, which provided extensive feedback and comments on the drafts; final approval of the cases was given by a business leader from each organization only after multiple rounds of review by the interviewees, their line managers and corporate communications representatives.

Qualitative Coding and Analysis

Throughout all three phases of the research, the researchers used an inductive grounded-theory approach to data analysis.[24] This approach involved both open and selective coding using NVivo software. We began by developing key categories to organize the data, closely adhering to participants' terms and language so as to preserve the intended meaning. Then, we distilled the categories down to more abstract (second-order) themes, derived research-based concepts from the literature, engaged in discussion within the research team and in iterative modeling that entailed abductive reasoning, and iterated between theory and evidence. Finally, we synthesized the second-order themes into the four dimensions of our AIX concept.

About the Authors

Ida Someth

Ida Someh (i.asadi@business.uq.edu) is a senior lecturer in the Business Information Systems discipline at The University of Queensland Business School, and a research affiliate at the Center for Information Systems Research (CISR), MIT Sloan School of Management. Her research investigates the organizational and societal impact of data, analytics and artificial intelligence. She completed an award-winning Ph.D. at The University of Melbourne and was the winner of the Paul Gray Award for the most thought-provoking article published in *Communications of the Association for Information Systems* in 2019.

Barbara H. Wixom

Barbara Wixom (bwixom@mit.edu) is a principal research scientist at the MIT Sloan Center for Information Systems Research (CISR). Since the mid-1990s, Barbara has deeply explored how organizations effectively generate value from information assets. She has published in journals such as *Information Systems Research*, *MIS Quarterly*, *MIS Quarterly Executive* and *Sloan Management Review*, and presents her work globally. In 2017, Barbara was awarded the Teradata University Network Hugh J. Watson Award for her contributions to the data and

24 One of the most prominent articles addressing how to conduct inductive qualitative data analysis is Gioia, D. A., Corley, K. G. and Hamilton, A. L. "Seeking Qualitative Rigor in Inductive Research," *Organizational Research Methods* (16:1), July 2012, pp. 15-31.

analytics academic community via the Teradata University Network.

Cynthia M. Beath

Cynthia Beath (Cynthia.Beath@mccombs.utexas.edu) is a professor emeritus of IS at The University of Texas at Austin. She received her M.B.A and Ph.D. degrees from UCLA and is an AIS LEO award winner. Her 2019 book, *Designed for Digital* (co-authored with Jeanne Ross and Martin Mocker) describes how organizations redesign themselves for the digital era. Cynthia has published in leading information systems research journals as well as *Harvard Business Review* and *Sloan Management Review*. An active advocate for her professional community, she hosted the IS field's first junior faculty consortium, has served on the AIS Council and co-founded *MIS Quarterly Executive*.

Angela Zutavern

Angela Zutavern (azutavern@alixpartners.com), managing director of AlixPartners, pioneered the application of machine intelligence to business leadership and strategy to drive meaningful insights, and has decades of experience in AI and other areas of digital technology, including data and analytics, platforms, privacy and governance. Angela helps clients to implement robust and sophisticated digital solutions to enable better decision making. She is a frequent speaker on the power of AI and the author of *The Mathematical Corporation: Where Machine Intelligence and Human Ingenuity Achieve the Impossible*. Angela is a member of the MIT Center for Information Systems Research's Data Research Advisory Board.