BUILDING AI EXPLANATION CAPABILITY FOR THE AI-POWERED ORGANIZATION

Ida A. Someh, *Senior Lecturer in Business Information Systems*
UQ Business School at The University of Queensland, Australia

Barbara H. Wixom, *Principal Research Scientist*
MIT Sloan Center for Information Systems Research (CISR)

Cynthia M. Beath, *Professor Emerita*
University of Texas, Austin

Organizations must strengthen and draw on five advanced data monetization capabilities—data science, data management, data platform, customer understanding, and acceptable data use—to deploy AI[1] solutions that are operationally, economically, and ethically sound.[2] Yet these five capabilities alone are insufficient for building stakeholders' confidence in an AI solution.[3] AI teams must also regularly engage with a variety of stakeholder groups to explain how the organization is managing characteristics of AI that impede trust in AI solutions. In this briefing, we present four challenging AI characteristics, and describe AI Explanation capability: an emerging enterprise capability that arises from practices AI teams use to build stakeholder confidence in AI solutions.[4]

FOUR AI CHARACTERISTICS THAT IMPEDE TRUST

Four characteristics of AI make it difficult for AI project teams to build stakeholder trust in AI solutions:

- **Unproven value:** Because the market still lacks a broad selection of proven AI use cases, leaders are uncertain if and how their company could create meaningful returns from AI project investments.
- **Model opacity:** AI draws upon computational mathematics and statistics that, due to their sophistication, make it very hard for even some data scientists (let alone laypeople) to readily understand if a model is producing accurate results in ways that are compliant and ethical.
- **Model drift:** An AI model will produce results that are biased in the same way as the data used to train it. Over time, as the real-world changes, or if error is introduced into data sources, the accuracy of the predictions an AI model produces from new data will deviate from the model's performance during the training period—a phenomenon known as "drift."
- **Mindless application:** AI models evaluate a new case and draw conclusions about it with some degree (or lack) of confidence. Because there is some level of error in transferring learning from historic case evaluations used for model training to new cases, AI model results are not definitive; applying the results as if they were definitive to new cases or in new contexts can be risky.

Our research found that organizations that adopt practices that resolve problems arising from these challenging AI characteristics successfully build stakeholder confidence.

- **Value formulation practices** tackle the problem of unproven value by helping AI teams substantiate, in terms that appeal to both direct and indirect stakeholders, exactly how AI can be an attractive investment.
- **Decision tracing practices** contend with model opacity by helping AI teams unravel model computations and mathematics, to convey the how behind model results to those who consume and make use of the output.
- **Bias remediation practices** enable AI teams to redirect model drift by exposing model decisions, helping the

1 We define artificial intelligence (AI) as applied analytics models that have some level of autonomy.

2 In the June 2022 MIT CISR research briefing, we described why and how organizations build the five advanced data monetization capabilities. See B. H. Wixom, I. A. Someh, and C. M. Beath, "Building Advanced Data Monetization Capabilities for the AI-Powered Organization," MIT Sloan CISR Research Briefing, Vol. XXII, No. 6, June 2022, https://cisr.mit.edu/publication/2022_0601_AdvancedAICapabilities_WixomSomehBeath.

3 B. H. Wixom, I. A. Someh, A. Zutavern, and C. M. Beath, "Explanation: A New Enterprise Data Monetization Capability for AI," MIT Sloan CISR Working Paper, No. 443, July 2020, https://cisr.mit.edu/publication/MIT_CISR-wp443_SucceedingArtificialIntelligence_WixomSomehZutavernBeath.

4 From Q3 2019 to Q2 2020, the researchers collected data on 52 AI projects 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 from 48 companies. The data was analyzed to identify practices that teams adopted in response to AI challenges and to inductively develop what we now call AIX capability. For more detail on and additional examples of building an AIX capability, see I. A. Someh, B. H. Wixom, C. M. Beath, and A. Zutavern, "Building an Artificial Intelligence Explanation Capability," *MIS Quarterly Executive* 21, No. 2 (June 2022): 143–163, https://aisel.aisnet.org/misqe/vol21/iss2/5/.

teams to reduce bias in a model's training activities throughout its life.

- **Boundary setting practices** address mindless application by guiding AI teams in applying AI model output mindfully to avoid unexpected outcomes and unintended consequences.

Collectively, these four sets of practices build an organization's AI Explanation (AIX) capability (see figure 1), which we define as *the ability to manage AI initiatives in ways that ensure models are value-generating, compliant, representative, and reliable*. As part of its AI journey, Microsoft adopted such practices to build an AIX capability. We draw on a case study about Microsoft's journey to illustrate examples of the practices the company has leveraged.[5]

AIX CAPABILITY PRACTICES AT MICROSOFT

In 2021, Microsoft was a $168 billion technology company headquartered in Redmond, Washington. The company employed 190,000 people who operated out of more than 600 office buildings worldwide. Microsoft's Real Estate and Facilities (RE&F) group was responsible for managing this extensive network of facilities. In 2015, RE&F approached Enterprise Data Science, Microsoft's internal-facing data science unit, to inquire if AI could be used to reduce the cost of managing these facilities.

AI Explanation (AIX) capability is the ability to manage AI initiatives in ways that ensure models are value-generating, compliant, representative, and reliable.

Microsoft Enterprise Data Science was the shared services group that helped internal Microsoft business units initiate and execute AI projects.[6] The group's objective was to learn from its internal engagements and, over time, diffuse what it had learned throughout the company in the form of better data science capability. Enterprise Data Science participated on the AI project team that developed an AI model for RE&F to predict building occupancy, then later helped to recontextualize the model for a variety of use cases for Microsoft and its customers. The group knew from experience that stakeholders would balk at supporting AI if they didn't fully understand it, and so began adopting practices to build AIX capability.

Value Formulation Practices

To address the unproven nature of AI, Microsoft's AI teams tracked AI value consistent with the goals of each case's distinct stakeholders. For example, for early projects championed by facilities leaders who managed space optimization, the AI project team monitored building occupancy, which surfaced opportunities for cost reductions. The AI model was later recontextualized to generate a dynamic HVAC schedule that automatically adjusted heating and cooling patterns of Microsoft building HVAC systems and produced significant cost savings. When possible, the project team also identified value for indirect stakeholders, such as that the optimized HVAC schedule increased employee comfort, which helped employees be more focused and productive.

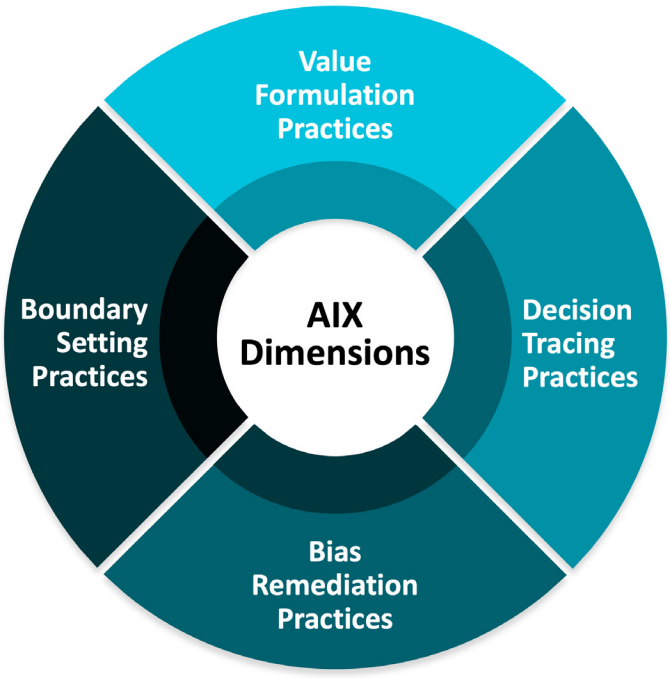
Practice examples:

- Tracking AI value
- Managing multi-stakeholder value propositions

Decision Tracing Practices

To address model opacity, the project team regularly described AI model mechanics to key stakeholders, using dia-

Figure 1: Four Sets of Practices that Build AI Explanation Capability



5 B. H. Wixom, I. A. Someh, R. M. Ionescu, and D. E. Leidner, "Microsoft's AI Scaling Journey for Occupancy Prediction," MIT Sloan CISR Working Paper No. 453, February 2022, https://cisr.mit.edu/publication/MIT_CISRwp453_Microsoft_WixomSomehIonescuLeidner.

6 I. A. Someh and B. H. Wixom, "Microsoft Turns to Data to Drive Business Success," MIT Sloan CISR Working Paper No. 419, July 2017, https://cisr.mit.edu/publication/MIT_CISRwp419_MicrosoftDataServices_SomehWixom.

grams and charts to lay bare and explain the model dynamics in play. The team also worked with domain experts to scrutinize the evidence behind model predictions and adapt model logic accordingly. These practices educated both the stakeholders and the AI team regarding if and how a model was producing accurate results.

Practice examples:

- Describing the model's inner workings to stakeholders
- Communicating the model's logic to stakeholders using visualizations
- Involving domain experts in adapting model logic

Bias Remediation Practices

Addressing model drift involved exposing model outputs to stakeholders and identifying potential bias. When the project team discovered an unusual pattern, stakeholders would review the pattern to help make sense of it. In one case, when results from the AI model suggested that a building was vacant, the stakeholders confirmed it, responding that a group of workers had just moved out of the building. To prevent rare building-use patterns from distorting model results, the AI project team reviewed datasets and excluded exceptions from the training data sets. The team also audited model results by comparing them with results obtained from other investigations. For example, as a part of an unrelated effort, another company physically surveyed a Microsoft office building to determine occupancy of open spaces. Data scientists from the AI project team were able to validate some of the team's occupancy model by comparing outcomes from the model with findings from the physical occupancy survey.

Practice examples:

- Involving domain experts in reviewing and justifying the AI model's output
- Examining data for bias potential
- Auditing the AI model's output
- Comparing human- and model-generated results

Boundary Setting Practices

Addressing AI's mindlessness required establishing clear boundaries for the model and scoping the model-based actions accordingly. Many of the early occupancy projects were associated with buildings on Microsoft's main campus at the company's headquarters in Redmond. Over time, the project team shifted focus to include other geographies, and discovered that each geography had its own distinct occupancy needs. As the team examined the limitations of the model, they learned that adapting the model to new contexts required new data collection and model training efforts.

Practice examples:

- Establishing boundaries for a model regarding where and when it performs well
- Examining model reliability in different contexts or scenarios
- Capturing new data and retraining the model as cases or contexts shift

Over time, these four types of practices helped Microsoft's Enterprise Data Science group mature the company's AIX capability and build stakeholder trust in AI solutions.

BUILDING AN AIX CAPABILITY FROM AIX PRACTICES

The AIX capability is multidimensional. It includes the ability to articulate a model's value for different stakeholders, to make the inner workings of models understandable, to create explanatory interfaces to expose and enable rectification of biases in AI model outputs, and to set boundaries for the safe application of AI models.

To establish AIX as an enterprise capability, it's crucial that an organization surface and share good practices, which takes effort and leadership.

Unlike the five data monetization capabilities, which are well understood and formalized, practices that build AIX capabilities are still emerging. In fact, most AI project teams are creating their playbook as they go. To establish AIX as an enterprise capability, it's crucial that an organization surface and share good practices, which takes effort and leadership. New, effective AIX practices should be seen as building blocks for an enterprise AIX capability. With this capability in hand, projects can stop crafting practices from scratch.

Therefore, organizational leaders must seek to proactively develop an AIX capability. A good starting point is to identify units and other organizations that are already driving effective explanations. Next, identify practices that the organization's own AI project teams have invented or adopted to create AI models that are value-generating, compliant, representative, and reliable, and enable sharing and reuse of these practices across the company. Finally, continue testing the most promising practices, and institutionalize the best ones. These actions will go a long way toward making AI trustworthy and consumable in an AI-powered organization.

MIT SLOAN CENTER FOR INFORMATION SYSTEMS RESEARCH

Founded in 1974 and grounded in the MIT tradition of rigorous field-based research, MIT CISR helps executives meet the challenge of leading dynamic, global, and information-intensive organizations. We provide the CIO and other digital leaders with insights on topics such as business complexity, data monetization, and the digital workplace. Through research, teaching, and events, the center stimulates interaction among scholars, students, and practitioners. More than seventy five firms sponsor our work and participate in our consortium.

CISR RESEARCH PATRONS

AlixPartners
Avanade
Axway, Inc.
Collibra
Pegasystems Inc.
PricewaterhouseCoopers
Standard Bank Group (South Africa)
The Ogilvy Group

CISR SPONSORS

Allstate Insurance Company
Amcor
ANZ Banking Group (Australia)
Australian Taxation Office
AustralianSuper
Banco Bradesco S.A. (Brazil)
Banco do Brasil S.A.
Bank of Queensland (Australia)
BlueScope Steel (Australia)
BNP Paribas (France)
Bristol-Myers Squibb
Cabot Corporation
CarMax
Caterpillar, Inc.
CEMEX (Mexico)
Charles River Laboratories, Inc.
CIBC (Canada)
Cochlear Limited (Australia)
Commonwealth Superannuation Corp. (Australia)
Credit Suisse (Switzerland)
Cuscal Limited (Australia)
CVS Health
DBS Bank Ltd. (Singapore)
Doosan Corporation (Korea)
Fidelity Investments
Fomento Economico Mexicano, S.A.B., de C.V.
Fortum (Finland)
General Mills, Inc.
General Motors Corporation
Henkel AG & Co. (Germany)
Hitachi, Ltd. (Japan)
HSBC Technology & Services (USA) Inc.
Johnson & Johnson (J&J)
Kaiser Permanente
King & Wood Mallesons (Australia)
Koç Holding (Turkey)
Mercer
National Australia Bank Ltd.
Nomura Holdings, Inc. (Japan)
Nomura Research Institute, Ltd. Systems Consulting Division (Japan)
OECD
Pacific Life Insurance Company
Pioneer Natural Resources USA Inc.
Posten Norge AS (Norway)
Principal Financial Group
Procter & Gamble
QBE
Raytheon Technologies
Reserve Bank of Australia
Santander UK/Grupo Santander
SC Global Tubular Solutions
Scentre Group Limited (Australia)
Schneider Electric Industries SAS (France)
Scotts Miracle-Gro
State Street Corp.
Stockland (Australia)
Suncorp Group (Australia)
Teck Resources Limited (Canada)
Tetra Pak (Sweden)
Trinity Health
Truist Financial Corporation
UniSuper Management Pty Ltd (Australia)
USAA
Webster Bank, N.A.
Westpac Banking Corporation (Australia)
WestRock Company
Wolters Kluwer
Zoetis Services LLC

Information as of July 2022

MIT CISR is funded by Research Patrons and Sponsors, and we gratefully acknowledge their financial support and their many contributions to our work.

Sponsorship and benefits: cisr.mit.edu/community/sponsor-and-patron-benefits

MIT CISR research publications: cisr.mit.edu/research-library

MIT Sloan School of Management
Center for Information Systems Research

245 First Street, E94-15th Floor
Cambridge, MA 02142
t 617-253-2348 | e cisr@mit.edu

cisr.mit.edu |