MIS Quarterly Executive

Volume 24 | Issue 2 Article 3

June 2025

The Danish Business Authority's Approach to the Ongoing Evaluation of AI Systems

Oliver Krancher

Per Rådberg Nagbøl

Oliver Müller

MIS Quarterly Executive

The Danish Business Authority's Approach to the Ongoing Evaluation of AI Systems

AI systems that work as intended when first deployed may drift over time, making increasingly inaccurate or biased decisions. Organizations therefore need to ensure the proper ongoing functioning of their AI systems, especially as the world around these AI systems changes. In this article, we describe the strategies used by the Danish Business Authority, an early public-sector adopter of AI, to ensure the effective ongoing evaluation of AI systems and provide four recommendations for other organizations.[1,2]

Oliver Krancher
IT University of Copenhagen (Denmark)

Per Rådberg Nagbøl
IT University of Copenhagen (Denmark)

Oliver Müller
Paderborn University (Germany)

Why AI Evaluation Is an Ongoing Challenge

Organizations are heavily investing in AI technologies such as machine learning (ML) to automate business processes and improve decision-making. Notwithstanding the enormous potential benefits of AI, its use may produce harm, such as when organizations make wrong or biased AI-assisted decisions. For example, the Dutch Tax Authority used an AI-based fraud detection system that was biased against people with a non-Western background, causing depression, burnout, divorces and suicides among the victims of the system and resulting in fines of over €6 million ($6.5 million)[3] for the agency and the resignation of the Dutch government.[4]

With growing awareness of AI's dark side and increasing pressure to comply with legislation such as the European Union's (EU's) AI Act, organizations have begun to set up structures to ensure the responsible use of AI. For example, organizations are striving to make the logic and outputs of AI systems more explainable,[5] undertaking rigorous evaluation before putting

1 Hind Benbya is the accepting senior editor for this article.
2 We thank the Danish Business Authority for its support of this research project.
3 Currency conversion rate as of March 2025.
4 Heikkilä, M. *Dutch Scandal Serves as a Warning for Europe Over Risks of Using Algorithms*. POLITICO, March 29, 2022, available at https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.
5 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 2024, pp. 259-278.

DOI: 10.17705/2msqe.00113

AI systems into production,[6] putting "humans in the loop" and "enveloping" AI systems by constraining inputs and outputs.[7]

Though these mechanisms contribute to a more responsible use of AI in organizations, they are not enough. In our research, we observed the case of an AI system that accurately recognized signatures on a document when the system was first deployed but performed this task less accurately a few months later when citizens started using new digital signature technologies. In other cases, people affected by an AI system's decisions may figure out how it works and try to game it by changing their inputs accordingly. Regardless of whether the accuracy of an AI system's decisions decreases due to external events (such as new digital signature technologies) or due to users adapting their interactions with the system, it is clear that AI systems working as intended at the time of their first deployment may degrade in performance as time goes by and the world around them changes.

For businesses and authorities wishing to harness the potential of AI, this means that evaluating AI systems is not a one-off task, finished when a system goes live; instead, it is an ongoing requirement.[8] Organizations therefore need solid capabilities for the ongoing evaluation and improvement of AI systems so they can detect when these systems drift and need to be retrained, redeveloped or decommissioned. Without these capabilities, organizations risk their AI systems making flawed and biased decisions that not only jeopardize the long-term economic returns from AI projects but may also violate legal requirements, such as post-market monitoring and evaluation mandated by the EU's AI Act.[9]

Unfortunately, there is little guidance on this issue in the literature. As with other types of digital innovations, the development and implementation of AI systems attract more attention from both managers and researchers than their ongoing maintenance. However, it is well known that the majority of the costs associated with IT projects accrue during maintenance and that actions during maintenance are critical to the beneficial use of IT systems.[10] For AI systems, part of the problem is that most organizations have only recently started their AI journeys, thus limiting their opportunities to learn from their own and other organizations' experiences.

Purpose and Focus of this Article

The purpose of this article is to provide organizations with advice on the ongoing evaluation of AI systems. This advice is based on the Danish Business Authority's (DBA's) experiences in evaluating its 14 operational AI systems, which have been in use for several years, making the DBA an AI pioneer in the public sector. As a government organization in the EU, the DBA has always been driven by a strong obligation to use AI responsibly. These efforts have been recognized by the EU and Gartner.[11] The pioneering role of the DBA provides a

6 See, for example, the AI risk assessment tool proposed in Nagbøl, P. R., Müller, O. and Krancher, O. "Designing a Risk Assessment Tool for Artificial Intelligence Systems," *Proceedings of the 16th International Conference on Design Science Research in Information Systems and Technology*, in Chandra Kruse, L., Seidel, S., and Hausvik, G. I. (eds) The Next Wave of Sociotechnical Design, Springer, 2021.

7 The term "enveloping" draws an analogy to robotics, where the actions and areas of movement of a robot are constrained (enveloped) to avoid harm. For a case study of how this concept can be applied to the responsible use of ML, see Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. "Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems," *Journal of the Association for Information Systems* (22:2), March 2021, pp. 325-352.

8 For a study of 55 AI implementation projects that identified dynamic environments and the ensuing need for continuous monitoring as one of five major challenges, see Hopf, K., Müller, O., Shollo, A. and Thiess, T. "Organizational Implementation of AI: Craft and Mechanical Work," *California Management Review* (66:1), October 2023, pp. 23-47.

9 For an introduction to the AI Act and how organizations may respond to it, see: 1) Blackman, R. and Vasiliu-Feltes, I. "The EU's AI Act and How Companies Can Achieve Compliance," *Harvard Business Review*, February 2024, available at https://hbr.org/2024/02/the-eus-ai-act-and-how-companies-can-achieve-compliance; and Renieris, E. M., Kiron, D. and Mills S. "Organizations Face Challenges in Timely Compliance With the EU AI Act," *MIT Sloan Management Review*, June 2024, available at https://sloanreview.mit.edu/article/organizations-face-challenges-in-timely-compliance-with-the-eu-ai-act/.

10 For an in-depth discussion of "our obsession with the new," see Vinsel, L. and Russell, A. L. *The Innovation Delusion: How Our Obsession with the New Has Disrupted the Work That Matters Most*, Currency, 2020.

11 See: 1) *AI Watch. European landscape on the use of Artificial Intelligence by the Public Sector*, European Commission, June 2022, available at https://ai-watch.ec.europa.eu/publications/ai-watch-european-landscape-use-artificial-intelligence-public-sector_en; and 2) *Case Study: How to Apply Ethical Principles to AI Models (Danish Business Authority)*, Gartner, August 2021, available at: https://www.gartner.com/en/documents/4004387.

unique setting to examine structures, challenges and solutions in the ongoing evaluation of AI systems. In a six-year research collaboration, we worked with the DBA to develop and implement mechanisms for the ongoing evaluation of AI systems. The insights reported in this article are derived from this collaboration and 16 interviews specifically focused on evaluating in-production AI systems (see the Appendix for further details about our research method). First, though, we provide a brief overview of the EU AI Act.

The EU AI ACT Requires Evaluation of AI Systems Throughout their Lifecycle

As a government organization inside the EU, the DBA is fully committed to the EU AI Act and aims to play a pioneering role in the responsible use of AI in governments. The EU AI Act marks a significant regulatory effort to ensure that AI systems developed and used inside the EU are safe and transparent and respect fundamental rights. The act includes a risk-based framework for categorizing AI systems based on their potential impact, ranging from minimal risk (e.g., ML-based spam filters) to unacceptable risk (e.g., ML-based systems for social scoring). Though AI systems in the latter category will be banned inside the EU, high-risk AI systems, such as those used for providing public services, are subject to stringent requirements.

The EU legislation has profound implications for the governance, development and use of AI systems. It mandates that organizations incorporate compliance considerations throughout the entire AI lifecycle, from development to deployment and maintenance. Operational procedures must be in place to meet requirements regarding, among others, risk assessment, transparency and human oversight. For example, before deployment, organizations must identify and mitigate potential risks of deploying AI systems, provide technical documentation about the systems and inform end users that they are interacting with an AI system.

After deployment, the act requires organizations to continuously evaluate the accuracy of predictions or decisions made by an AI system, monitor the system's robustness in terms of performance degradation, discrimination and cyberattacks, and keep records of user interactions with the system for potential future investigations. Because these activities occur after deployment, these activities are called "post-market monitoring" in the EU AI Act.

Data governance is another critical area affected by the EU AI Act. The act mandates that high-risk AI systems must be developed using high-quality, relevant and representative datasets that should be as free of errors and biases as possible. Moreover, organizations are required to provide information about how their AI systems function, explain predictions or decisions made by AI and ensure that humans can intervene when necessary.

In essence, the EU AI Act makes compliance pivotal in developing, deploying and using AI. It establishes the need to evaluate AI systems holistically throughout their complete lifecycle. Compliance with the act will mean that organizations avoid legal repercussions and contribute to the responsible advancement of AI technologies.

Next, we provide an overview of the DBA and its AI activities and then describe the strategies it implemented in post-market monitoring, specifically the ongoing evaluation of AI systems.

Overview of the Danish Business Authority and Its AI Activities

The Danish Business Authority (DBA) is a government agency within the Danish Ministry of Industrial, Business and Financial Affairs. With approximately 1,200 employees, the DBA's mission is to create the "best conditions and framework for responsible Danish companies" and to "make it easy and attractive to run a responsible business and create development throughout Denmark." The core focus of the DBA is to ensure the growth of the Danish economy through business development, effective business regulation and enforcement and by providing businesses with professional and efficient services. The DBA operates a cross-governmental IT platform with over 20 million yearly visitors, oversees more than 800,000 registered companies submitting over 300,000 annual reports a year, receives over 500,000

Table 1: Selected AI Systems at the Danish Business Authority

AI System	Description
Signature	Assesses whether company-founding documents are signed
COVID Compensation	Predicts the likelihood of fraud in businesses applying for the reimbursement of their fixed costs during the COVID-19 pandemic
Company Registration	Predicts the likelihood of fraud in newly registered Danish companies
Auditor's Statement	Predicts the likelihood that the valuations of a company's assets in an auditor statement are correct and the statement does not contain violations
Sector Code	Supports the verification of a company's industry sector code

Table 2: Digital Tools in the MLOps Platform Supporting the AI Lifecycle at the Danish Business Authority

Tool Name	Tool Category	Description
Record Keeper	Data and pipeline versioning tool	Supports the tracking of data and models used for training and evaluation
Race Track	Model deployment and serving tool	Supports the automated deployment of AI systems to Kubernetes infrastructure
Catwalk	Dataset labeling and annotation tool	Supports domain experts in labeling data for training or evaluation
Control Tower	Model monitoring tool	Provides feedback on caseworkers' ongoing use of AI models

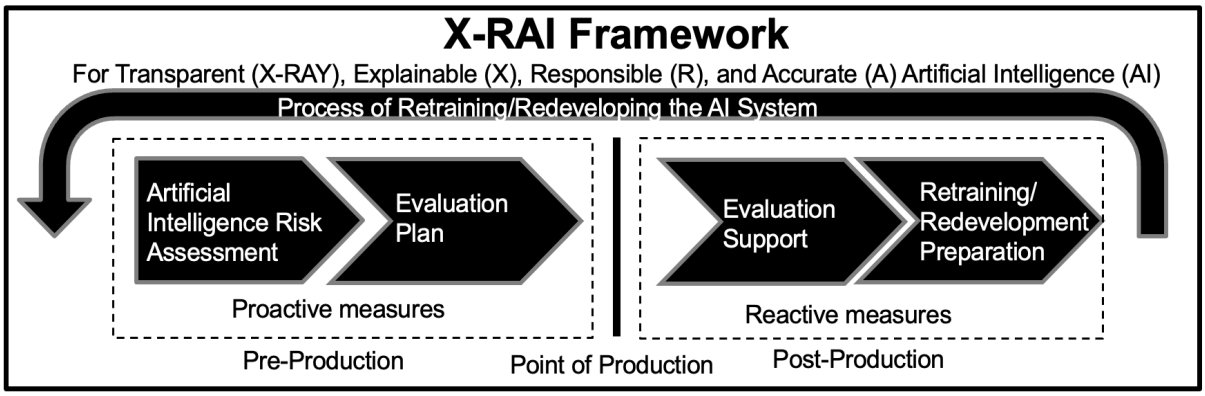
phone calls to its support centers and makes 5.7 million automated decisions per year (i.e., decisions where businesses or citizens receive an instantaneous response rather than awaiting human review and approval). For example, the compensation schemes for businesses suffering from the COVID-19 pandemic included more than 400,000 online applications and 250,000 final settlements and automated payouts, totaling 55 billion Danish Crowns ($7.9 billion). This high case workload and the high degree of digitalization in the Danish public sector present a fertile environment for the use of AI.

In 2017, the DBA began to experiment with building in-house machine learning (ML) models, eventually establishing the ML Lab. Seven years later, the ML Lab consisted of a data science manager and a data engineering manager, a designer and facilitator of compliance and governance (who is a co-author of this article), a user experience designer, a product owner, three data engineers and eight data scientists. Fundamental principles for using AI at the DBA include in-house know-how and resources, a business-driven approach to AI, high AI ethics standards, explainability and decision support, simplicity over complexity, and not mindlessly following the hype. These principles mean that the ML Lab takes care of the entire AI infrastructure with close collaboration between users and domain and technical experts throughout the entire lifecycle of AI systems, from early development to shutdown or redeployment of the AI systems. Table 1 provides brief descriptions of the five AI systems mentioned in this article (a subset of the 14 operational AI systems).

In addition to the 14 operational AI systems, the DBA also has a technical MLOps Platform, called the "Intelligent Control Platform," which comprises infrastructure and a set of tools to support the management of AI systems across their lifecycle. Table 2 provides an overview of these tools. At the heart of the infrastructure is a NEO4J graph database management system[12] with over 450 million nodes. The tools include

12 For information on NEO4J, see *Neo4j AuraDB: Fully Managed Graph Database*, available at https://neo4j.com/product/auradb/.

Figure 1: X-RAI Framework for Responsible Use of AI at the DBA



Race Track for deploying AI systems, Catwalk for storing AI system outputs and evaluating AI systems through data annotation, and Record Keeper for tracking configurations and events. In addition to these digital tools, the DBA uses X-RAI, a framework for responsible AI use, which is described next.

X-RAI Framework for Ongoing Evaluation of AI Systems

In our six-year research collaboration with the DBA we developed the X-RAI framework for AI governance to ensure the responsible use of AI and ensure compliance with the EU AI Act. X-RAI is an acronym for transparent (X-ray), explainable (X), responsible (R), accurate (A) artificial intelligence (AI). As shown in Figure 1, X-RAI consists of four questionnaire-based sub-frameworks.[13] Though the framework addresses issues beyond the ongoing evaluation of AI systems, two of the four sub-frameworks—"evaluation plan" and "evaluation support"—guide evaluation activities. Moreover, the first sub-framework—"artificial intelligence risk assessment"—establishes important foundations for evaluating AI systems. Thus, the X-RAI framework forms the basis for the ongoing evaluation of the DBA's AI systems.

Artificial intelligence risk assessment (AIRA): AIRA assesses pre-production risks from the early stage of development to the production-ready project. It involves stakeholders from different backgrounds, including domain experts and data scientists. The AIRA questionnaire includes topics such as performance metrics, consequences of correct and incorrect predictions, explainability, fairness,

13 The questionnaires comprise a set of questions for data scientists and domain experts. Following the principle of structured intuition, these questions help ensure that these actors consider and make informed choices about issues relevant to the risk and evaluation of AI systems. The questionnaires can be downloaded from https://pure.itu.dk/ws/portalfiles/portal/106272075/X-RAI-_A_Framework_for_a_Transparent_Explainable_Responsible_and_Accurate_Use_of_Artificial_Intelligence.pdf. For more details of the research underpinning the framework, see: 1) Nagbøl, P. R., Müller, O. and Krancher, O. "Designing a risk assessment tool for artificial intelligence systems," *Proceedings of the 16th International Conference on Design Science Research in Information Systems and Technology*, in Chandra Kruse, L., Seidel, S., and Hausvik, G. I. (eds.) *The Next Wave of Sociotechnical Design*, Springer, 2021; and 2) Nagbøl, P. R. *Theoretical and Practical Approaches to the Responsible Management of Artificial Intelligence*, Ph.D. Thesis, IT University of Copenhagen, 2022, available at https://pure.itu.dk/ws/files/99933954/PhD_Thesis_Final_version_Per_R_dberg_Nagb_l.pdf.

Table 3: The DBA's Strategies for the Ongoing Evaluation of AI Systems

Stage	Strategies
1. Pre-Production	1.1. Envelop evaluations
	1.2. Decide when to evaluate
	1.3. Establish traceability
	1.4. Automate deployments
2. Production	2.1. Exercise human oversight
	2.2. Promote reflective use
	2.3. Revise job profiles
	2.4. Monitor performance over time
3. Formal Evaluations	3.1. Conduct formal evaluations with deliberate sampling and "blind" AI output
	3.2. Build tools to ease evaluation
	3.3. Leverage synergies

personal data and data quality. The information collected through AIRA is not only crucial for pre-production evaluation but also provides an essential foundation for evaluating AI systems throughout their lifecycle.

Evaluation plan: Before an AI system is put into production, the evaluation plan guides managers, domain experts and data scientists on agreeing how, when and by whom the AI system will be evaluated post-implementation. Answering the evaluation plan questions before an AI system goes live helps ensure that post-implementation evaluations are planned and the required resources are available.

Evaluation support: The evaluation support sub-framework supports the ongoing evaluation of an AI system in production, ensuring that it meets the quality criteria. Evaluation support structures and documents the ongoing evaluations, including making decisions on whether the AI system should continue in production or be decommissioned, retrained or redeveloped.

Retraining/redevelopment preparation: The retraining/redevelopment preparation sub-framework guides domain experts, data scientists and managers in planning the process of retraining or redeveloping an AI system, including identifying dependencies with other AI systems and the need for training data.[14] Once an AI system is scheduled for retraining or redevelopment, the AIRA and evaluation plan sub-frameworks must be revisited, as indicated by the cyclical arrow in Figure 1.

Strategies for Effective Ongoing Evaluation of AI Systems

As mentioned above, both the EU and Gartner have highlighted the DBA's AI governance efforts as exemplary. Despite its extensive use of AI systems, the DBA is not aware of a case where AI has caused harm or has been subject to public debate. This suggests that other organizations can learn from the DBA's AI governance efforts, including its effective ongoing evaluation of AI systems. We therefore now describe the strategies that the DBA has adopted at the three stages of an AI systems lifecycle—pre-production, production and post-implementation formal evaluations—to ensure the effective ongoing evaluation of AI systems (the strategies are summarized in Table 3). Though some of the strategies are executed during the formal post-implementation evaluation of AI systems (Stage

14 The DBA uses the term redevelopment for changes to AI systems beyond retraining, including changing the algorithm or the features of an AI system.

3), effective ongoing evaluation also hinges on strategic decisions and preparations made pre-production (Stage 1) and on strategies adopted during production—i.e., actions taken before formal evaluations while an AI system is in use (Stage 2).

Stage 1. Pre-Production Strategies

At the pre-production stage, the risks associated with an AI system are assessed before deploying it to production. This stage is also crucial for establishing the foundations for the evaluation of AI systems once they are in use. The DBA has adopted four key Stage 1 strategies.

1. Envelop evaluations: A fundamental principle at the DBA is to "envelop" AI systems—i.e., to set boundaries for using an AI system, akin to building a virtual wall that sets a boundary for a vacuum cleaning robot.[15] But the DBA realized it is also important to envelop AI evaluations—i.e., to set clear boundaries on how an AI system designed today should be evaluated tomorrow. The envelopment of evaluations involves defining who will be responsible for evaluating an AI system and based on what data they use, which metrics they will use and the targets for those metrics. The DBA ensures the envelopment of AI systems by defining the acceptable use of a given system in the artificial intelligence risk assessment component of the X-RAI framework, while it ensures the envelopment of evaluations by prescribing evaluation modalities in the evaluation plan component.

According to a DBA manager, enveloping AI systems and their evaluations helps "tighten the bolts [and] sharpen the tools." Envelopment prevents users from deviating from the intended use of the system and from the agreed-upon modalities of evaluating it. This manager metaphorically compared not enveloping AI systems to a construction site, where workers "remove envelopment from dangerous situations just because it makes something easier for them"—i.e., where removing protection mechanisms would ease completing an urgent task. Similarly, business units might be tempted to expedite their work by moving beyond the agreed-upon use of an AI system or by continuing to use it despite its lack of evaluation. Specifying and enforcing acceptable use and evaluation modalities up-front through the AI risk assessment and evaluation plan components of the X-RAI framework are important strategies for avoiding such behaviors.

2. Decide when to evaluate: An important issue for ensuring effective ongoing evaluations of AI systems is defining when to evaluate. The DBA found it was important to specify the timing of evaluations in the evaluation plan before putting AI systems into production. These planning efforts helped ensure that evaluations received sufficient priority while also making effective use of resources, given that formal evaluations were a laborious task.

Critical considerations for the timing of evaluations include the patterns of contextual change in the domain of the AI system and the availability of resources. Based on these considerations, the DBA chooses between event-based, frequency-based and seasonal evaluation (see Table 4).

Event-based evaluations are triggered by events that impact the AI system's performance or change the context so that the predictions are no longer suitable. Examples of such events include changes in technical standards and the changes of industrial classification codes used in the DBA's Sector Code AI system.

Frequency-driven evaluations are driven by events recurring after a fixed timespan—for example, every third month or every year. Our interviewees highlighted several factors that determine the evaluation frequency. One issue is how long the organization can run on a false premise, given that the pattern in behavior could, in principle, change the day after the evaluation. A DBA data scientist identified a question to ask when planning a frequency-driven evaluation: "For how long can we tolerate that the AI system answers incorrectly and live without discovering a drop in AI system performance?"

Other factors affecting the frequency of evaluations mentioned by our interviewees were the impact of the AI system, the thoroughness of prior evaluations and the extent of dynamics. For example, one domain expert stated:

> *"[Other factors are] how big an impact it has [and] how the earlier assessments*

15 For a detailed description of the DBA's strategies for enveloping AI systems, see Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. op. cit., December 2024.

Table 4: Scheduling Options for Ongoing AI Evaluations

Timing	Description	Examples	Suitable Situations
Event-Based	An evaluation is scheduled near an event in the domain of the AI system.	Changes in industry sector codes trigger evaluation. Changes in AI system that provides input to a focal AI system.	AI systems with predictable changes in their domain or in related systems
Frequency-Based	An evaluation is regularly scheduled after a specific period.	Evaluations are conducted once a year.	AI systems not suitable for event-based or seasonal evaluation logic
Seasonal	Evaluations are scheduled at specific times of the year.	Evaluations are scheduled to avoid yearly workload peak times.	AI systems with seasonal fluctuations in case workload

> *... looked. ... If we had the first evaluation rather quickly and then held another one after three months, and everything looks fine, and it is business as usual, then there is no reason we should meet again in three months. Then we can set it up to be biannual."*

Seasonal evaluations are appropriate for activities that fluctuate depending on the time of the year or other recurrences. For example, the DBA uses the Auditor's Statement AI system to support caseworkers in assessing companies' annual reports. Companies submit their reports to three deadlines a year. The DBA schedules evaluations deliberately to avoid these peak times.

3. Establish Traceability: The DBA recognized the need to establish not only management and planning foundations but also technological foundations for the ongoing evaluation of AI systems. Over the past years, the DBA has built the MLOps-based Intelligent Control Platform, comprising a set of tools and infrastructure to support the management of AI systems across their lifecycle (see Table 2 above). A key component of this platform is Record Keeper, which, according to a DBA manager, helps track precisely "what the model was at a certain date, what the state of that model and the composition of that model was." This traceability capability is achieved by strongly relying on infrastructure as code,[16] policy as code,[17] versioning systems and integration with graph databases, helping to reproduce the conditions under which a given model was trained and evaluated. Without knowing these conditions, it would be difficult if not impossible for the DBA to discover if the input data of an AI system in production is comparable to the data on which the model was trained or evaluated. Another benefit of Record Keeper is that it helps maintain an overview of when an AI system has been evaluated.

4. Automate deployments: A second technological foundation for effective ongoing evaluation of AI systems is fully automated AI system deployment. A DBA data scientist told us that its Race Track digital tool automatically deploys AI systems "within a few minutes and, depending on the size of a model, up to one hour." A DBA manager emphasized that automated rapid deployment is essential for keeping control of the AI systems in use: "If we discover that a model diverges from the precision we need, we need to replace that model. The longer we wait with the replacement, the more erroneous data it will put out." He also said that automated deployment processes help ensure the integrity and consistency of AI systems: "We have deliberately decided that applications in

16 Infrastructure as code means managing computing resources through machine-readable files rather than through manual configuration or physical hardware changes.

17 Policy as code means defining and enforcing security rules through machine-readable files.

production cannot be [changed] manually. They only exist through automated processes. So, even a developer with heightened rights cannot change the production environment manually. This … guarantees that the model has not been changed by human hands after it came into production." Together, Record Keeper and Race Track give the DBA full control of the configuration of AI systems in use, enabling it to reproduce AI systems exactly, evaluate them when needed and rapidly rectify issues.

Stage 2. Production Strategies

While the evaluation plan prescribes and schedules formal evaluations to rigorously assess AI systems in use, formal evaluations are conducted only at specific points in time, often once a year. To ensure prompt feedback about the functioning of AI systems, the DBA adopted four strategies during the production stage, described by our interviewees as the period between an AI system going live and its first formal evaluation.

1. Exercise human oversight: A key strategy for the responsible use of AI at the DBA, as in many other organizations, is human oversight—i.e., putting humans "in the loop" for all decisions that can impact people or businesses. In the words of a DBA domain expert: "We never start a case [against a business] based on the AI system's prediction unless a human has made a professional assessment of the case. We never automatically initiate anything just because a computer has said so."

Human oversight allows humans to detect deterioration in the quality of AI systems. For example, when new digital signature tools entered the market, the DBA's Signature AI system, designed to detect whether applications are signed, produced an increasing number of false positives (i.e., applications that the Signature system classified as lacking a signature even though they were signed.)[18] The caseworkers processing these false positive cases recognized the pattern of false positives and reached out to the DBA's ML Lab to retrain the model with new signatures.

2. Promote reflective use: The efficacy of human oversight in gathering ongoing feedback about the functioning of AI systems can be increased by adopting reflective use practices. Such practices have various facets. Some interviewees highlighted team reflection meetings, in which complex cases were discussed, as an opportunity to assess how well the AI system diagnosed the cases. A DBA data scientist told us: "In our Tuesday meetings, we have a 10-minute evaluation session. If there are five people and each person processes 10 cases a week, then we can quickly talk about 50 cases. This way, you can have a quick evaluation once a week."

Another facet of reflective use is the shared culture of viewing AI as a tool, recognizing that AI systems may not produce objective truth and may be biased due to the training data. A DBA domain expert emphasized that reflective use requires a critical mindset from users: "ML is just a tool that is no better than the person who feeds the model. You can throw a lot of features into the model … but if I, as a caseworker, have a bias against a specific sector, then the model will be somewhat… subjective."

This critical mindset is essential for recognizing problems in AI systems. Closely related to viewing AI as a tool is a feeling of collective responsibility for ensuring the responsible use of AI. A DBA manager said that X-RAI has been instrumental in shaping this "strong sense of responsibility … which is a core part of [the DBA's] identity."

3. Revise job profiles: The extensive use of AI systems means that evaluation has become a new part of DBA caseworkers' job profiles. Highlighting evaluation in caseworkers' formal job profiles helps them to recognize the importance of evaluation work and ensures that they give evaluation sufficient priority. As mentioned by a DBA manager, it also helps ensure that users have the skills required for the critical use of AI systems during production: "In one team, we can see very clearly that they are profiling themselves differently. In their job postings, they mention machine learning and data deduction all over the place."

4. Monitor performance over time: Another strategy adopted by the DBA for ensuring feedback about the functioning of AI systems was to develop systems that facilitate monitoring of AI system performance over time. Though the

18 The Signature AI system is designed to detect the absence of a signature, so a positive case is where the signature is absent. Thus, a false positive denotes a case that the Signature AI system incorrectly flagged as unsigned, despite being signed.

DBA was experimenting with more sophisticated "drift detection" solutions at the time of our data collection, it had developed the "Control Tower" digital tool, which provides the DBA's ML Lab with feedback about the functioning of the AI systems. The Control Tower, located at the interface between AI models and case handling systems, allows caseworkers to manually "mute" certain features of an AI model. Muting a parameter means ignoring its impact on the model's output. When caseworkers mute certain features repeatedly, this may indicate a problem in the AI model and, hence, the need to evaluate the model.

Stage 3. Formal Evaluation Strategies

Although strategies used during the production stage, such as human-in-the-loop, provide critical feedback on the functioning of AI systems, they are insufficient for ensuring that the AI systems behave as they should. The DBA's core mechanism for maintaining ongoing control of AI systems is formal evaluations. The DBA has adopted the following three formal evaluation strategies:

1. Conduct formal evaluations with deliberate sampling and "blind" AI outputs: Relying solely on human-in-the-loop as a strategy for monitoring whether AI systems behave as they should would pose two critical problems with the DBA. The first is *sampling bias*. At the DBA, human caseworkers focus mainly on positive cases identified by the AI systems, such as businesses likely to commit fraud. Indeed, the principle of human oversight mandates that humans assess an AI-positive case before taking action against a business or a citizen. Conversely, caseworkers look less frequently at AI-negative cases (e.g., cases with a low fraud score). As a consequence, the chances of humans detecting an AI system drifting toward an increasing number of false negatives—i.e., cases wrongly classified as negative by an AI system—are relatively low.

The second problem is *self-reinforcing loops of bias*. During production, caseworkers can see the AI system's predictions (e.g., a fraud score) and may be influenced by it, as emphasized by a data scientist: "If you give a caseworker a percentage score, say it's 98% likely that this company commits money laundering, then chances are high that the caseworker will look at the case much more thoroughly than one where the percentage score is 5%."

Because caseworkers are likely influenced by the AI system's predictions, any bias in these predictions is likely to cause a bias in caseworkers' decisions. If these decisions are taken as "ground truth" during evaluation, the bias will not be recognized. If the bias is not recognized and eliminated, there is a risk that humans will, over time, come to believe in the wrong relationships in the AI system. As a consequence, it is less and less likely they will detect the bias.

To address these two issues, the evaluation plan and the evaluation support sub-frameworks guide the DBA in conducting formal evaluations with deliberate sampling and blind AI outputs. Formal evaluations can mitigate the limitations of adopting a purely human oversight strategy in three ways. First, formal evaluations allow a more comprehensive assessment of cases identified as negative by the AI system. For example, a data scientist described the use of formal evaluations in the COVID Compensation AI system: "We used the ML model to score all cases. Then we took the 100 cases with the highest fraud score plus 100 randomly drawn cases from the rest of the population (not including the top 100), shuffled them and sent them to the business for evaluation." Thus, by including not only the most likely positive cases but also a random sample, the team ensured a stronger representation of negative cases, substantially increasing the chance of identifying false negatives (i.e., cases that the AI system erroneously classified as negative).

Second, because humans may be biased by the AI systems' recommendations, formal evaluation presents the opportunity for blinding the AI system's output—i.e., allowing caseworkers to assess a case without seeing the AI system's output. As a data scientist explained: "If you use another caseworker who is not the one who originally looked at the case, the assessment would not be biased [by the AI system's prediction]."

Third, formal evaluations provide opportunities for systematic bias assessments, as explained by a domain expert through a fictitious example: "If you sit with a number of cases and the ML model only picks foreigners, then you

might think this cannot be correct. … Then, it would be great to integrate these variables, such as residency, into your dataset and test some hypotheses [about the effect of residency on the decisions of the ML model]."

Blind evaluations based on deliberate, partially random, sampling are the ideal at the DBA but are not always feasible or needed. Blind evaluations require that caseworkers label a case without seeing the decision that the AI system proposed or a caseworker previously made. Blind evaluations were feasible in the COVID Compensation AI system, but in other systems, such as in the Company Registration AI system, it is impossible to see the information relevant to a case without also seeing the history of decisions made for a given case, including the decision made by a caseworker based on AI output. Thus, blind evaluations require that the information is presented in such a way that caseworkers cannot see the decision made on a case or any subsequent events that are a consequence of this decision.

Blind evaluations may not be needed in other contexts. For example, a DBA manager said that labeling data for evaluating the Signature AI system is "very, very easy" because it just involves assessing whether there is a signature on a document. In this case, the risk that a caseworker may be biased by knowing an AI system's decision beforehand is minimal.

2. Build tools to ease evaluation: The DBA developed an infrastructure not only to increase traceability, automate deployments and observe performance over time but also to make evaluations less burdensome. For example, the Catwalk user interface allows evaluators to annotate relevant data in a system that automatically stores annotation data and makes it accessible for retraining purposes, which helps to ease resource bottlenecks. Both domain experts and data scientists emphasized the advantages of such a digital infrastructure. According to a domain expert, Catwalk also makes it easier to pass on qualitative feedback to data scientists: "If we make comments [in Catwalk], the comments will continuously be sent to our ML teams, so you can ask them to reconsider the model on the basis of your comments."

3. Leverage synergies: Though the DBA sometimes found it difficult to ensure sufficient resources for evaluation, our interviewees shared that several strategies help create synergies between evaluation and other activities, which may help to address resource issues and reduce tedious elements of evaluation work. Specifically, they recommended looking for synergies between regular work, AI system evaluation, human training and AI system training. For example, a domain expert pointed out potential synergies between evaluations and training of new hires: "We have just hired a new employee on the team who needs training. We always focus them on [the AI system] because there are some good cases for training."

Another source of synergies is to store labeled evaluation data and recycle it as training data when retraining the AI system. This is why the retraining/redevelopment preparation sub-framework prescribes an analysis of whether evaluation data can be reused for retraining (though this does not obviate the need for creating another dataset to evaluate the retrained system).

The DBA has also explored the idea of using data from human oversight for formal evaluations in AI systems where the risk of biasing caseworkers through AI recommendations is minimal, such as in the Signature AI system. A domain expert told us: "… the best in the world would be that our case management is constructed in a way if I, for example, had processed a case in the Signature [AI system], then I could, while closing my handling of the case, do some evaluations of the positives."

Recommendations for Applying the DBA's Strategies in Other Organizations

In the previous sections, we have described the strategies adopted by the DBA for ensuring the effective ongoing evaluation of AI systems. Drawing on the learnings from the DBA case, we now provide four recommendations for other organizations.

1. Establish a Framework for the Ongoing Governance of AI Systems

We recommend that organizations implementing AI systems should establish a robust post-market AI governance structure

to ensure ongoing evaluation and control over these systems. Managers should build this governance framework by assigning clear roles and responsibilities, particularly for the regular monitoring and evaluation of AI performance. These responsibilities include estimating the time and resources needed for ongoing evaluations and defining a structured schedule based on the anticipated risks and rewards of each evaluation. AI governance should also mandate the definition of specific metrics for assessing key dimensions of AI performance, such as accuracy, fairness, robustness and potential bias. By setting clear metrics for these and other relevant criteria, managers will enable structured tracking and comparison of AI performance over time. This governance structure would provide decision makers with insights into AI performance trends, providing a data-driven basis for timely decisions about model updates, retraining or even decommissioning systems when necessary.

2. Establish Practices for Fostering Reflective Use

To build a culture of reflective use among teams interacting with AI systems, we recommend that managers implement practices that encourage critical engagement with AI outputs. Reflective use involves actively questioning and assessing AI outputs rather than simply accepting them. Teams should arrange regular peer discussion sessions where team members critically examine recent cases or results flagged by AI systems. These discussions allow users to challenge AI decisions, consider alternative interpretations and reinforce the understanding that AI is just one of many decision-support tools rather than an omniscient authority. Additionally, managers could rotate tasks among team members or assign roles that expose employees to various aspects of AI implementation, such as data labeling, model assessment and user experience. By broadening team members' perspectives, organizations can help develop a more nuanced and critical understanding of AI capabilities. Offering AI training specifically focused on critical and ethical thinking can further enhance employees' ability to interpret AI outputs confidently and responsibly.

3. Build a Digital Infrastructure to Automate Monitoring and Enforce Evaluation

A critical component of AI governance is implementing a digital infrastructure that supports ongoing monitoring and evaluation. We recommend that managers invest in tools that log evaluations, automate data labeling and track performance drifts or anomalies in real time. Developers should incorporate "enforced evaluation" systems with automated stop functions that deactivate AI systems if they do not meet scheduled evaluation requirements. These systems ensure compliance with governance standards and provide early warning signals for potential issues. Additionally, managers should implement "AI control towers" or centralized dashboards that consolidate metrics from different AI systems and thus provide a comprehensive overview of AI health across the organization. By raising automated alerts when specific thresholds are reached, these digital tools would enable managers to proactively address issues before they affect broader operations, ensuring a consistent and effective evaluation cycle.

4. Embed Ongoing Evaluation into AI System Design

We recommend that, when designing new AI systems, organizations embed evaluation-friendly features to support long-term quality assurance. Managers should ensure that new AI systems are configured to facilitate feedback flows directly from end users and other stakeholders so that data scientists can receive ongoing, real-world insights into the AI system's performance. Design choices should include options to conduct blind evaluations, reduce bias in feedback and enable random sample injections that help detect hidden errors. Adopting these design principles will enable organizations to improve system quality while dynamically monitoring the AI system's real-world impact.

Concluding Comments

While many businesses and public-sector organizations are experimenting with AI technologies, their current focus is on identifying suitable use cases for AI, developing AI systems to

address these use cases and assessing whether it would be responsible to deploy these systems to production. But as more and more AI solutions come into everyday use, organizations will need to redirect their attention from development to operations and maintenance and to how they can set up structures to ensure that their growing portfolio of AI systems behaves as it should in an ever-changing world. Some organizations may implement pre-production risk assessments and use humans to ensure the responsible ongoing use of AI systems. Other organizations from highly regulated industries, such as banking and pharmaceuticals, are used to auditing their systems regularly and will be comfortable with applying their existing approaches to AI systems.

However, our research collaboration with the DBA shows that, though pre-production risk assessments, humans in the loop and regular system audits may be helpful, they are not sufficient to ensure the responsible ongoing use of AI systems. The challenge is that AI systems that work as they should when first deployed may make less accurate or more biased decisions as the world around these systems changes or as systems are retrained. Due to the inscrutability of AI and the complexity of the domains in which these systems are embedded, it is very difficult to identify when an AI system falls short of expectations. In the worst case, AI systems in use are like blindfolded drivers, uncertain whether the road ahead has changed since they last saw it and whether the vehicle is still on track.

We have described how the DBA addressed the challenge of the ongoing and effective evaluation of AI systems by combining several strategies, each of which has limitations that can be compensated for by the strengths of other strategies. For example, humans in the loop can produce continuous insights into the performance of an AI system, allowing the organization to recognize drift quickly. However, humans may be biased by the outputs of the AI system they oversee; they may not detect false negatives in their everyday work and may not detect any problems at all if they unquestioningly trust the AI system. On the other hand, mandatory regular evaluations mitigate these issues by blinding the AI system's output and ensuring coverage of negative cases. The downside of regular, mandatory evaluations is that they do not produce continuous insights into model performance and require a significant amount of relatively tedious labeling work from caseworkers who might be busy with other tasks.

Given that a weakness of one strategy is often a strength of another, it is clear that organizations need to adopt multiple complementary strategies to ensure the effective ongoing evaluation of AI systems. By adopting the X-RAI framework developed at and used by the DBA, organizations can identify strategies that can help them jumpstart their efforts of establishing practices and processes for the ongoing evaluation of AI systems.

Appendix: Research Method

The findings in this article are derived from a six-year collaborative research program we undertook at the Danish Business Authority (DBA). Most of the research program was based on action design research, which focuses on building solutions, such as the X-RAI framework, for practice-inspired research problems, such as AI evaluation, through iterations of building, intervention and evaluation. The empirical material collected during the research came from 16 semi-structured qualitative interviews conducted over three rounds with data scientists, domain experts, managers and legal experts at the DBA. The interview questions in the first two rounds focused on challenges in the ongoing evaluation of AI systems and the strategies used or proposed for addressing these challenges. The questions in the last round focused on ascertaining the extent to which and under what conditions the DBA has implemented the strategies. The interviews were transcribed and analyzed using an inductive coding approach similar to coding techniques in the grounded theory method that aggregate openly coded material to higher-order categories emerging from the data. Quotes from the interviews, some of which have been translated from Danish to English, are included in the article.

About the Authors

Oliver Krancher

Oliver Krancher (olik@itu.dk) is an associate professor in the Digital Business Innovation

Section of the IT University of Copenhagen. His research focuses on digital sourcing, automation and project management. He has published in leading information systems and software engineering journals, including *Journal of Management Information Systems*, *Journal of the Association for Information Systems*, *Journal of Strategic Information Systems*, *Journal of Information Technology*, *European Journal of Information Systems* and *Empirical Software Engineering*. Oliver has received several awards for his work, including the McKinsey Business Technology Award and the SIG Sourcing Early and Mid-Career Award.

Per Rådberg Nagbøl

During this research, Per Rådberg Nagbøl (Per.Nagbol@outlook.com) was at the IT University of Copenhagen, where he earned his Ph.D. and participated in a co-funded research collaboration with the Danish Business Authority. He is now a senior data and AI governance professional at Novo Nordisk. He has used action design research to design systems and procedures for AI governance. His work has been published in *Journal of the Association for Information Systems* and *MIS Quarterly Executive*. He received the Best Student Authored Paper Award at the 2021 Design Science Research in Information Systems and Technology Conference.

Oliver Müller

Oliver Müller (oliver.mueller@uni-paderborn.de) is a professor of management information systems and data analytics at Paderborn University, Germany. His research focuses on data-driven decision-making, including the design and effective use of innovative ML solutions for supporting human decision-making and the acceptance and implications of data-driven decisions in organizations. His research has been published in *Journal of Management Information Systems*, *Journal of the Association of Information Systems*, *Journal of Strategic Information Systems*, *European Journal of Information Systems* and elsewhere. Oliver has received several awards and honors for his work, including AIS Best Information Systems Publication awards in 2017 and 2023.