MIS Quarterly Executive

Volume 19 | Issue 4 Article 7

December 2020

Challenges of Explaining the Behavior of Black-Box AI Systems

Aleksandre Asatiani

Pekka Malo

Per Rådberg Nagbøl

Esko Penttinen

Tapani Rinta-Kahila

See next page for additional authors

Challenges of Explaining the Behavior of Black-Box AI Systems

Authors

Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara

MIS Quarterly Executive

Challenges of Explaining the Behavior of Black-Box AI Systems

There are many examples of problems resulting from inscrutable AI systems, so there is a growing need to be able to explain how such systems produce their outputs. Drawing on a case study at the Danish Business Authority, we provide a framework and recommendations for addressing the many challenges of explaining the behavior of black-box AI systems. Our findings will enable organizations to successfully develop and deploy AI systems without causing legal or ethical problems.[1,2]

Aleksandre Asatiani
University of Gothenburg (Sweden)

Pekka Malo
Aalto University School of Business (Finland)

Per Rådberg Nagbøl
IT University of Copenhagen (Denmark)

Esko Penttinen
Aalto University School of Business (Finland)

Tapani Rinta-Kahila
The University of Queensland (Australia)

Antti Salovaara
Aalto University School of Arts, Design and Architecture (Finland)

Organizations Need to Be Able to Explain the Behavior of Black-Box AI Systems

Huge increases in computing capacity and data volumes have spurred the development of applications that use artificial intelligence (AI), a technology that is being implemented for increasingly complex tasks, from playing Go to screening for cancer. Private and public businesses and organizations are deploying AI applications to process vast quantities of data and support decision making. These applications can help to reduce the costs of providing various services, deliver new services and improve the safety and reliability of operations.

However, unlike conventional information systems, the algorithms embedded in AI applications can be "black boxes." Previously, those who developed applications could completely explain how an algorithm worked. Given an input, they could tell you what the output would be and why, because the systems applied human-made rules. That is no longer true for AI-based applications. The application creates internal structures that determine outputs, but these are inscrutable to outside observers, and even the programmers cannot tell you why a specific output was generated. Many AI systems leverage machine learning,

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 insightful feedback that has greatly improved the quality of this article. We are grateful to the Danish Business Authority for sharing their time and allowing us to conduct this study.

DOI: 10.17705/2msqe.00037

where a model learns how to act by detecting patterns in data by employing only general principles for how such patterns can be found. The actual process of finding those patterns may remain hidden and there is no human input or intervention in the process.

As a consequence, information systems (IS) researchers are striving to find ways to improve the transparency of algorithms embedded in AI applications—i.e., to provide the ability to explain the rationale or logic behind algorithmic decisions to human stakeholders. IS researchers and academics refer to this area as the "explainability"[3] of black-box AI algorithms.

The ability to explain how AI algorithms reach their decisions is a legal requirement in Europe. The European Union's General Data Protection Regulation (GDPR) mandates an individual's right to explanation. From an ethical point of view, the ability to explain can help to identify and defuse problematic biases. For example, Amazon's face-recognition and recruitment models were found to develop racial and gender biases.[4] Similarly, from a safety perspective, the ability to explain can help to identify the source of the problem in cases where an AI application has—from the users' point of view—made a mistake. Explanations can help to prevent such problems from reoccurring.

Thus, in their search for greater performance, organizations must deploy AI applications in a legal, ethical and safe manner, which means they must have the ability to explain how the applications make their decisions. This is especially true in the public sector, where public trust and confidence in AI-based decisions are of paramount importance.

Although there have been several attempts to produce technical explanations that allow humans to understand the behavior of AI applications, this is not always feasible because of the inductive reasoning applied by many AI applications. Technology giants (including Google and IBM) are beginning to offer AI solutions that are, at best, partially explainable and, in the U.S., the Defense Advanced Research Projects Agency has a program dedicated to the task of developing explainable AI. An inability to provide sufficient and meaningful explanations creates barriers for the successful deployment of AI applications in an organization, and therefore hinders the potential benefits of higher operational efficiency and accuracy.

Explaining the behavior of AI systems requires more than purely technical measures. Organizations must also consider what the outputs from the systems mean for human stakeholders.[5] A recent report[6] on using AI to combat public-sector fraud suggests that *"where a technical explanation for an AI tool is not possible, practical or meaningful, an ability to explain the priorities or strategic basis for a decision may suffice and may even be more meaningful ... depending upon the context."* Acquiring the ability to explain thus requires a managerial solution; however, there is a scarcity of such solutions.

Our research therefore addressed the question: How can organizations reconcile the growing demands for explanations of how AI-based algorithmic decisions are made with their desire to leverage AI to maximize business performance? This article presents the findings of our research, which are based on a case study of the Machine Learning Lab at the Danish Business Authority. (Details of the study are provided in Appendix A.)

First, we describe the six elements of a hypothetical intelligent AI agent—the model, goals, training data, input data, output data and environment. We then present a framework with six dimensions, each corresponding with one of the elements that will enable organizations to explain how AI-based algorithms reach their decisions. We then illustrate how this framework helped our case organization, the Machine Learning Lab at the Danish Business Authority, to responsibly and successfully exploit apparently unexplainable black-box AI. The lessons from this case are valuable both for IS researchers and

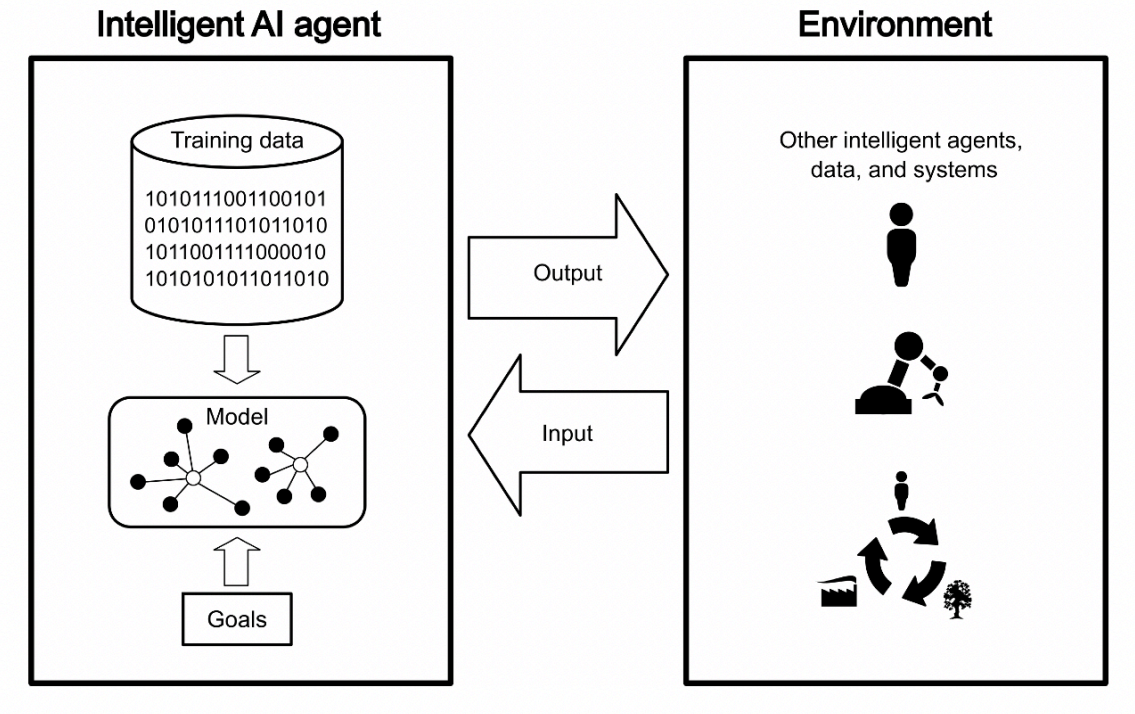
3 For a description of explainability, see Rosenfeld, A. and Richardson, A. "Explainability in Human–Agent Systems," Autonomous Agents and Multi-Agent Systems (33), May 2019, pp. 673-705.

4 See, for example, Vincent, J. "Gender and Racial Bias found in Amazon's Facial Recognition Technology (Again)," *The Verge*, January 25, 2019, available at https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender.

5 For more information, see Martin, K. "Designing Ethical Algorithms," *MIS Quarterly Executive* (18:2), May 2019, pp. 129-142.

6 The Use of Artificial Intelligence to Combat Public Sector Fraud: Professional Guidance, Serious Fraud Office [U.K.], in collaboration with New Zealand's Serious Fraud Office, February 2020. This report was prepared by members of the International Public Sector Fraud Forum.

Figure 1: The Six Elements of an Intelligent AI Agent



designers of black-box AI applications, and for organizations that deploy such applications.

The article concludes with four recommendations derived from our analysis of the case study. These recommendations provide executives with a toolbox for proactively managing issues concerned with explaining how AI algorithms work and thus helping them to reap the potential benefits of AI applications.

The Six Elements of an Intelligent AI Agent

Our framework is described by reference to a hypothetical autonomous intelligent AI agent (which is depicted in Figure 1). According to Russel and Norvig, an intelligent agent, whether human or machine, pursues goals by processing data and interacting with other agents in the environment.[7] The intelligent AI agent we reference has six main elements: the model, goals, training data, input data, output data and environment. These elements form the dimensions of our framework.

Typically, developers of a machine-learning-based application construct the AI model by defining its mathematical formulation, setting goals for it, and training it to reach those goals. The model, which is, in essence, a mathematical function that relates an input to an output, has parameters whose originally unknown values are specified via a suitable training algorithm. Obviously, the choice of model has implications for understanding and being able to explain how the AI agent works: a human may be able to follow the if-then paths of simpler models, but this might not be possible with more complex models.

The *goals* of an intelligent AI agent are performance metrics (such as accuracy levels or average prediction error rate) that allow developers and other stakeholders to evaluate

7 For more on this definition of an intelligent agent, see Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd edition, Prentice Hall, 2010.

whether the agent is satisfying the performance criteria set for it. By scrutinizing the agent's goals, people may be able to explain its behavior (for instance, testing the agent's output against various performance metrics might reveal imbalances in the original training data).

To prepare the intelligent AI agent for use, the algorithm is run on a set of *training data*, which the algorithm uses to identify suitable values for the model's as-yet-unspecified parameters. Supervised or unsupervised learning approaches may be used, depending on the business problem the AI agent is being used for.

In supervised learning, the AI agent is trained from a labeled dataset with data organized into predefined categories. For example, if the AI agent is to make predictions of a company's success or failure, the training data might include figures from annual financial statements that could be expected to predict business success. Once trained, a set of test data not used in its training is input to the AI agent. The agent's performance can then be evaluated against its goals (e.g., its ability to predict business risks accurately).

In contrast to supervised learning, unsupervised learning makes sense of the input data independently; it does not make use of neatly categorized training data. With unsupervised learning, the AI agent searches for hidden patterns (e.g., uncovering sources of business failures from combinations of financial and/or other indicators). In most applications, unsupervised learning is best described as an exploratory or descriptive tool.

Regardless of whether the learning approach is supervised or unsupervised, the body of data used to train the AI agent shapes its capabilities and is therefore integral to the model used. Some explanations of the behavior of the AI agent are rooted in biases found in the training data, which, for example, may reveal why the agent discriminates for or against certain groups of people.

Once the AI agent has been trained and validated, it is deployed for real-world use. Actual *input data* (e.g., figures from companies' annual statements) is fed into the black-box algorithm, which then produces *output data* (e.g., a probability of a business failing). Examination of the input and output data can reveal explanations for the AI agent's behavior. For instance, imprecise recording of input data may point to why there are flaws in the agent's output data. Comparing the output data to other available information can also help in tracing the agent's decision logic and finding blind spots in its operations.

The final element of the AI agent is the *environment* in which it operates. The environment determines the sources and validity of the incoming data, and the agent influences the environment via its outputs (e.g., the resultant risk assessment of a company's future may shape the actions of the company). Such feedback loops are especially important in "reinforcement learning," where the AI agent learns from interacting with its environment by trial and error and receives rewards for good performance. If the AI agent is deployed in a different environment, it is unlikely to operate correctly (e.g., a system trained to identify business risks may not perform well in non-business settings). Thus, an AI agent's inappropriate behavior might be explained by it being deployed in an environment for which it was not trained.

A Framework for Explaining the Behavior of Black-Box AI Systems

The above discussion suggests that the ability to explain the behavior of an AI agent can be enhanced by examining and suitably designing each of the six elements. Thus, as summarized in Table 1, our framework for explaining the behavior of black-box AI systems has six dimensions, each of which corresponds to one of the elements of the hypothetical AI agent.

Dimension 1: The AI System's Model

A core element of the ability to explain how an AI system operates is a thorough understanding of the model used—specifically, how it turns inputs into outputs. At the technical level, gaining this understanding can be fairly easy for simple, rule-based systems or certain machine-learning models such as decision trees and regressions. However, technical explanations may not be practical or even possible with more complex models where logical decision rules cannot be extracted, such as deep neural networks (layered computing systems whose structure resembles

Table 1: Six-Dimension Framework for Explaining the Performance of AI Systems

Dimension	Description	Example
1. Model	Explanation of the AI system's logic/behavior based on tracing its decision-making patterns.	A specific business-risk probability may be explained by the if-then sequence of steps taken by a business-risk estimation model.
2. Goals	Explanation of the AI system's logic/behavior derived from priorities or the strategic basis for a given decision.	The agent flags high probabilities of risk for companies that engage in reputation-compromising activities such as producing health-harming products or causing environmental damage, with the explanation lying in the fact that the model is trained and tested with performance metrics that give great weight to risking the organization's reputation.
3. Training Data	Explanation based on the characteristics of the training data.	The agent assigns exceptionally high probabilities of risk to certain types of business, such as medical practices, because of biased training data. Data on medical practitioners might have been collected in economically deprived areas while data from other businesses are geographically more diverse.
4. Input Data	Explanation based on the characteristics of the input data.	Unreliable business-risk probabilities can be explained by low-quality input data produced by inaccurate measurement of relevant risk factors.
5. Output Data	Explanation derived from humans' examination and verification of the output.	A human examines the validity of the AI agent's business-risk probability for a loan application and makes sure that the rationale for the decision can be explained to the applicant in meaningful terms.
6. Environment	Explanation that is based on the environment in which the AI agent operates.	Inappropriate risk estimations may be explained by the AI agent being fed risk-assessment data from environments that are not suitable for this purpose (e.g., using soccer-league scoring data to predict the risks of businesses not connected to soccer).

that of the biological networks of neurons in brains). Although the model's designers and developers most certainly understand the underlying mathematical formulation of their models, even they may find it very difficult, if not impossible, to explain the model's behavior once it has been trained and is used to process actual data.

The difficulty of providing a technical explanation is compounded in models where not only millions of parameters are learned from training data but the underlying structure or model topology is adjusted automatically by the training algorithm. The inability to explain how such an AI application has made a decision has caused problems in high-profile contexts, such as police trying to detect potential offenders before they have committed a crime.[8] Models that have been trained using unsupervised learning are typically more difficult to explain than supervised ones because of the lack of a priori labeling and benchmarking standards. Although reinforcement learning models can be assessed

8 See, for example, "Rules Urgently Needed to Oversee Police Use of Data and AI – Report," *The Guardian*, February 23, 2020, available at https://www.theguardian.com/uk-news/2020/feb/23/rules-urgently-needed-oversee-police-use-data-ai-report.

against various criteria, their trial-and-error-based learning logic makes them particularly challenging to explain.

Other dimensions of the framework can be used to explain the behavior of a black-box AI system. Whether these explanations are sufficient depends on several factors, including legislation, the impact of the decisions on stakeholders and ethical issues.

Dimension 2: The AI System's Goals

In setting goals for an AI system, developers need to translate high-level business objectives into concrete performance metrics that can be used to steer the development of the agent. Well-chosen metrics serve as the primary means for comparing the performance of competing inscrutable models against each other. Ideally, they can also help to explain a model's behavior by revealing situations where it performs well and where it fails. Consider, for instance, the example mentioned above of using a neural network for detecting potential offenders: If this AI application successfully identifies a high proportion of future offenders, it is deemed to have high accuracy. However, it might still produce an unacceptably large number of false positives within some groups (e.g., certain ethnic groups may be overrepresented) while failing to predict actual offenders in other groups, because the accuracy metric does not account for imbalances in the distribution of ethnicity data. Testing the AI system against a performance metric that does address this possibility helps to reveal such imbalances and, thus, provides explanations for the underlying logic. Such testing shifts the emphasis to training data, as discussed next.

Dimension 3: Training Data

The way in which an AI system performs is determined by the characteristics of the data used to train it. A biased training dataset leads to biased decisions even if there is nothing wrong with the functionality of the algorithm taught by the data. A good example is the AI system used in U.S. courts to predict convicts' risk of recidivism, which was found to have a racial bias.[9] The data used to train an AI system tends to reflect biases in the real world, which causes the AI system to adopt the same biases and therefore produce biased outputs. Even when the algorithm is too complex to be explained meaningfully, awareness of the characteristics of the training data can shed light on how and why it translates the input data into outputs in the way it does.

Dimension 4: Input Data

Insufficient attention to the quality of input data can result in difficulties in explaining the behavior of an AI system, as illustrated by a scandal in Australia. A simple AI system was deployed for identifying social-welfare debt and initiating debt collection from citizens.[10] Poor input data quality, stemming from pairing two incompatible data sources, caused the AI system's debt calculations to be incorrect. Although the algorithm was technically explainable, neither government workers nor citizens had been informed of the incompatibility of the sources the AI system was drawing on. Their impression was that the system was a black box, which made it difficult for them to prove the incorrectness of the debt calculations. As a consequence, workers and affected citizens suffered significant stress. A thorough understanding of the limitations resulting from matching incompatible datasets would have mitigated the problems that ensued.

Dimension 5: Output Data

The problems with the Australian AI system were aggravated by a decision to remove human workers from the debt-collection loop: the lack of human oversight of the AI system's outputs enabled erroneous debt claims to be sent to citizens. Having humans check the outputs becomes all the more important with a black-box AI system that employs opaque decision-making logic. The Russian proverb "trust but verify" is very apt: even if the model performs well, it may need a human gatekeeper.[11] Although the algorithm itself may be opaque, scrutinizing the

9 See Buranyi, S. "Rise of the Racist Robots – How AI is Learning All our Worst Impulses," *The Guardian*, August 8, 2017, available at https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

10 Bajkowski, J. "Federal Court bins Robodebt's Defective Algorithm," *iTNews*, November 27, 2019, available at https://www.itnews.com.au/news/federal-court-bins-robodebts-defective-algorithm-534677.

11 See Desai, D. R. and Kroll, J. A. "Trust but Verify: A Guide to Algorithms and the Law," *Harvard Journal of Law & Technology* (31:1), Fall 2017.

viability of its output can help humans provide explanations that are sufficiently meaningful.

Dimension 6: Environment

Understanding the boundaries of the environment in which an AI system operates can help to explain its decisions, even when it is not possible to explain the workings of the underlying algorithm. The importance of defining and knowing the environmental boundaries for an AI system is illustrated by the well-publicized case of Amazon's Alexa operating beyond its intended use context by recording personal conversations and emailing them to another Alexa user.[12] Although Alexa's actions seemed inexplicable at first, approaching them from the perspective of environmental boundaries helps to explain what was going on: although the AI system "thought" it was operating in a particular environment (i.e., taking orders from its human owner), it was, in fact, receiving input data from a context in which it should not have been operating (a private conversation between two humans).

These examples quoted above for each of the six dimensions of our framework suggest that explanations of the behavior of an AI system should holistically take account of all six dimensions. This is precisely the approach adopted by the Machine Learning Lab at the Danish Business Authority (DBA), as described below. This case study identified novel tools for tackling the challenges of explaining how AI applications reach their decisions, even though the inner workings were not always entirely explainable. This approach has enabled the DBA to implement AI applications responsibly and legally.

Machine-Learning AI Applications at the Danish Business Authority

The Danish Business Authority is an agency within Denmark's Ministry of Industry, Business and Financial Affairs. It has approximately 700 employees, divided between the headquarters in Copenhagen and two satellite departments in Silkeborg and Nykøbing Falster. Its primary responsibility is to enhance opportunities for business growth in Denmark, but it also has specific regulatory obligations, such as fraud prevention and supervision of companies without imposing an unnecessary administrative burden on the Danish business community. One of the DBA's obligations is to maintain and apply laws such as Denmark's Companies Act, Financial Statements Act, Bookkeeping Act and Commercial Foundation Act.

To facilitate the activities associated with these obligations, the DBA operates a multi-agency online platform called Virk (https://virk.dk). Citizens can use Virk, for example, to establish or shut down business enterprises, handle various registrations and submit documents such as financial reports electronically. The online business register contains approximately 809,000 companies, with 812,000 registrations, and filings of 292,000 annual reports. Annual reports are submitted in two formats: PDF documents to be read by humans, and documents in structured data format XBRL (eXtensible Business Reporting Language) to be automatically machine-processed. The sheer volume of data presents the DBA with ample opportunities to pursue machine learning for such core tasks as supporting companies' legal compliance, checking annual reports for signs of fraud, and identifying companies early on their route to distress so that timely support can be given.

Because of the large data volumes involved, the DBA established its Machine Learning Lab in 2017 to implement machine-learning projects for greater efficiency and scalability. The lab's team leader and chief data scientist, "James,"[13] stated the following:

> *"We are, in essence, trying to use [machine learning] as a force multiplier for our colleagues performing the controls but also trying to lessen the manual workload and reserving the human decision making for the more creative or advanced tasks."*

The lab uses technologies such as Neo4j's platform, Docker and Python[14] for the development, application and support of machine-learning AI applications, rather than

12 See Warren, T. "Amazon Explains How Alexa Recorded a Private Conversation and Sent it to Another User," *The Verge*, May 24, 2018, available at https://www.theverge.com/2018/5/24/17391898/amazon-alexa-private-conversation-recording-explanation.

13 Pseudonyms are used for all informants to protect their identity.

14 For information about Neo4j and its products, see https://neo4j.com/company/.

commercial off-the-shelf solutions. The lab develops functional prototypes of machine-learning applications that are capable of solving business problems specified by case workers: "*We are focusing on meeting the information need that the business has,*" James explained. A DBA steering committee decides whether to move a prototype to production use. When the committee decides in favor of implementation, an external vendor then implements the machine-learning application for real-world deployment. "David," an Early Warning Europe[15] case worker, elaborated on the importance of this type of governance:

> *"It's really easy to end up on the front page of a tabloid newspaper. … This is why [we] make sure the model is only handed over from the partner organizations to stakeholders through a package of management consultancy training, capacity-building, documentation, all these support services, where we make sure that at least they know the logic of using it."*

At a higher level, the lab is engaged in a wider dialogue about the use of AI in government and was recently involved in the Danish National Strategy for AI, with a particular focus on the transparent application of AI in the public sector.[16]

Denmark, in general, and specifically the DBA, is considered to be at the forefront of e-government initiatives globally. According to a recent UN report,[17] Denmark is a world leader in e-government development. Within the EU, Denmark is ranked first for the provision of e-government services for businesses,[18] and it was also ranked fourth in the EU's Digital Economy and Society Index (DESI),[19] where it was listed among the leaders in digital public services. Furthermore, Europe's Digital Progress Report specifically highlighted the DBA's Virk portal, noting that roughly 96% of Danish businesses make use of Virk. However, the high level of digitization and digitalization driven by the DBA in Denmark has not been accompanied by adverse media comments about digital government experienced by other countries. For these reasons, we consider the DBA to be a legitimate source for best practice in organizational use of AI. Conducting a case study of the DBA's development and implementation of AI applications enabled us to learn from a well-performing organization in the field of government IT. Details of the case study are in Appendix A.

How the Danish Business Authority Applied the Framework

The DBA's approach to explaining the behavior of its AI applications is characterized by limiting the capacities of AI agents while still obtaining the desired outputs from the applications.[20] The approach took account of all six dimensions of the framework described above: the choice of AI model, the goals of the AI application, the training data, input and output data, and boundaries of the environment in which the AI application operates. The actions taken by the DBA in all of these dimensions are summarized in Figure 2.

The key benefit of holistically managing the six very different dimensions is gaining a better understanding of, and control over, the outputs from AI applications, which enables the organization to prevent or at least mitigate any undesired outcomes. By establishing and knowing the boundaries of an AI system's operation, the organization has a better understanding of the system's capacity to act. Within these boundaries, AI solutions can be harnessed to maximum advantage—even those with models that are seemingly inexplicable. Thus, instead of being

15 Early Warning Europe provides free, impartial and confidential counselling to companies in distress. For more information, see https://www.earlywarningeurope.eu/.

16 See National Strategy for Artificial Intelligence, 2019, available at https://eng.em.dk/media/13081/305755-gb-version_4k.pdf

17 E-Government Survey 2020: Digital Government in the Decade of Action for Sustainable Development, United Nations Department of Economic and Social Affairs, August 24, 2020, available at https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2020.

18 See eGovernment Benchmark 2019: trust in government is increasingly important for people, European Commission, October 18, 2019, available at https://ec.europa.eu/digital-single-market/en/news/egovernment-benchmark-2019-trust-government-increasingly-important-people.

19 The Digital Economy and Society Index (DESI), European Commission, 2019.

20 An example of limiting an AI system's capacity provided in Robbins, S. "AI and the Path to Envelopment: Knowledge As a First Step Towards the Responsible Regulation and Use of AI-Powered Machines," *AI & Society* (35), April 10, 2019, pp 391-400.

Figure 2: The DBA's Approach Took Account of all Six Dimensions of the Framework

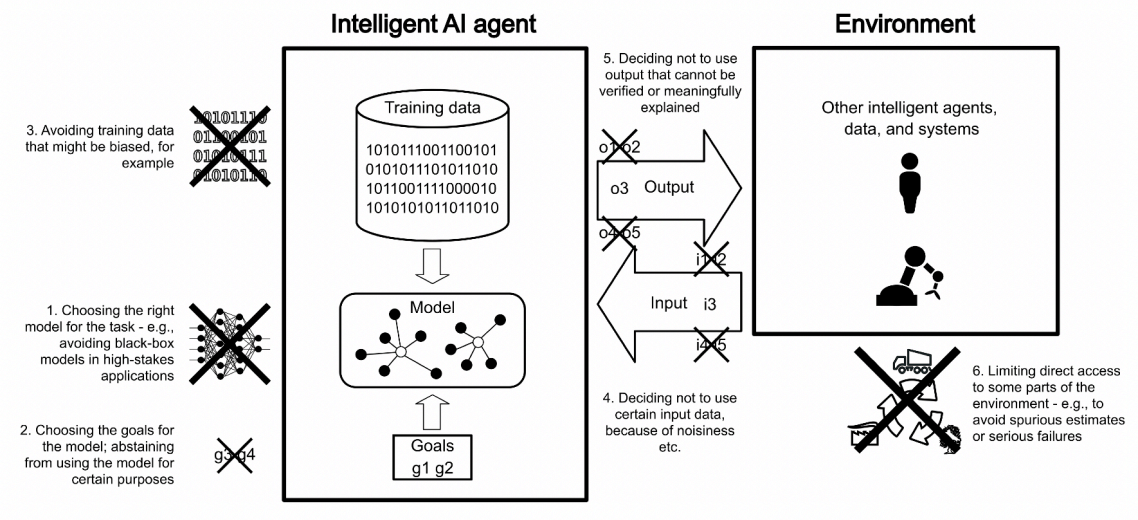


Table 2: Examples of AI Applications at the DBA

Project	Project Description	Goal	Input	Output	Model
Company Registration	To detect fraudulent behavior among newly registered Danish companies.	To prevent fraudulent companies from being established.	Data from the business registry, annual reports and VAT reports.	Probability of fraudulent behavior.	Gradient boosting (XGBoost).[21]
Signature	When coupled with its document filter, to speed up verification of whether company founding documents are signed or not.	To facilitate the process of founding a company.	Scanned images of the founding documents.	Probability of a document being signed or not.	Residual network (ResNet-16).[22]

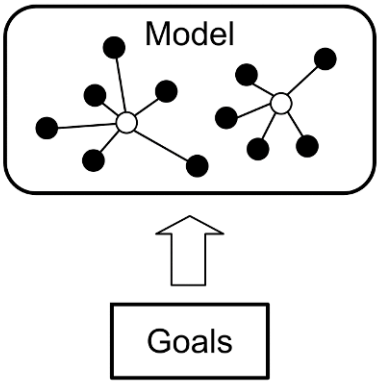
viewed as a method for producing technical explanations, the DBA's approach provides mechanisms for understanding and controlling the behavior of AI applications.

Below, we discuss in detail how the DBA's approach took account of each of the six dimensions. To demonstrate how the authority's actions were implemented, we provide examples from two ongoing projects at the DBA. These two projects, which exploit AI in different ways, are summarized in Table 2. (A full list of the AI projects being undertaken by the DBA is given in Appendix B.) The purpose of the first project, "Company Registration," is to prevent people from establishing companies for fraudulent purposes—i.e., creating companies that were never intended for the stated business objectives, but instead have ulterior, fraudulent motives behind them. In contrast, the aim of the second project, "Signature," is to facilitate the process of creating legitimate companies by detecting the absence of signatures from the documents

21 Gradient boosting is a machine-learning technique for regression and classification problems. XGBoost is an open-source software library for gradient boosting frameworks.

22 The residual-network technique uses machine-learning based on deep neural networks and is especially powerful in image-detection tasks. ResNet-16 is a residual-network technique whose architecture has 16 neural network layers.

Figure 3: Factors to Consider When Choosing and Controlling Training Data



- Consider how explainable the AI use case needs to be
- Select a model whose structure is not too open-ended, to avoid excessive flexibility that could allow learning from harmful spurious correlations
- Use structures that mirror the nature of the underlying problem

- Choose concrete and unambiguous performance metrics that reflect underlying business goals
- Analyze the pros and cons of the performance metrics carefully, and discuss the choices with various stakeholders

When contemplating the choice of model, ask:

- *Is this level of complexity necessary for achieving the required functionality?*
- *Could sufficient performance be obtained by means of a simpler alternative?*
- *Will the main users of the model need to explain its functioning to other people?*

required to found a company. Together, these two projects illustrate how the DBA's approach took account of all six dimensions of the framework for explaining the behavior of black-box AI systems.

Choosing the AI Model and Setting Goals (Dimensions 1 and 2)

To ensure that an AI application meets users' requirements, developers must carefully choose the system's model and goals (i.e., performance metrics), taking account of the need to be able to explain the outcome in a specific use case (see Figure 3).

Clearly, the demand for an explainable model depends on the type of project. For the Company Registration application, the DBA opted for an explainable model, because users must be able to understand readily why the algorithm has raised a red flag for a newly registered company:

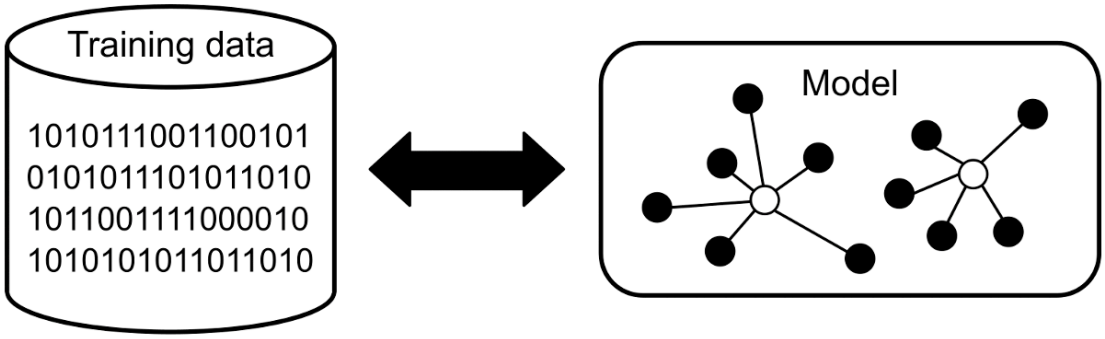
> *"We need to communicate the results and our findings to the case workers, so we try to use algorithms that are not complicated … or at least algorithms that can fairly easily give you some sense of which are the most important factors and which are not. So, [we need] explainable algorithms. … I guess that the more difficult it is for the case worker to actually see right away what the right answer is, the more important it is for the algorithm to be able to explain itself."* "Mark," a data scientist at the DBS Machine Learning Lab

The model chosen for an AI application has direct implications for how explainable the outputs from the application will be. Sometimes, though not always, choosing the model requires a tradeoff between performance and the transparency of the model selected. In most cases at the DBA, however, performance losses resulting from transparency demands have been negligible, as emphasized by James:

> *"We … [compared] a number of models, and gradient boosting came out as number one. We could have chosen deep learning [or] a deep neural network, but we chose not to, because we find [it would be] too complex to explain."*

In the Signature application, however, which is essentially an image-recognition application

Figure 4: Factors to Consider When Establishing Controls for Input and Output Data



- Gather an adequate set of training data
- Explore patterns and how they might influence the AI system's behavior
- Critically assess variables and their measurement
- Identify biases, and apply corrections if needed
- Prevent uncontrolled self-learning from potentially biased incoming data

When controlling training data, ask:

- *What kind of data exactly is needed for making the decisions?*
- *Are we using that sort of data, or something that is of lower quality and perhaps even biased, merely because it is more easily accessible?*
- *Is the data well suited to the type of AI model that we are using?*

using neural networks, it is perfectly acceptable to use a black-box approach. This is because users can easily verify whether the model works correctly or not without the need to understand its internal logic in great depth.

When choosing the goals and performance metrics for an AI application, our DBA interviewees emphasized that there is no silver bullet. Mark (a data scientist), reflected on how to prioritize among multiple performance metrics:

> *"... you could focus on the precision of the model. For example, how well does it predict [compared to our predictions] of ... fraudulent behavior in the future? How many would be correctly classified? But if [the case workers] have enough time on their hands, it might be [worthwhile looking retrospectively to] see how many of the companies [predicted to commit fraud] actually [do]. But that would give probably more work to the case workers. ... It depends on the situation, and it's a dialogue with the case workers exactly [as to] which metrics are the most important ones in each case."*

Clearly, the successful choice of metrics is highly problem-specific and requires both thorough understanding of the nature of the underlying data and solid domain expertise.

Understanding and Controlling the Training Data (Dimension 3)

Training data plays a key role in determining how an AI application works once deployed (see Figure 4). At the DBA, managers were well aware of the need for high-quality training data: "*It is my head on the line if it seems that the data is not good enough or [the data] is biased"* ("Steven," a data scientist at the Machine Learning Lab).

Controlling training data requires access to sufficient quantities of data and in-depth knowledge of the data, including any inherent biases and limitations. For example, if training data is limited to smaller companies, the implications of this bias should be assessed and the model's applicability may be narrower than initially assumed (the AI application might be suitable only for smaller firms). To ensure the application is relevant for companies of all sizes, any such biases in the training data will need to be corrected. To guarantee high-quality training data for both the Signature and Company Registration AI applications, the DBA opted to tag a large body of data manually, using domain experts as consultants in this process. In the words of Steven:

> *"We had tagged data, we had around 6,000 tagged documents, so we had a pile of [documents] that had not been used in training or in developing, so we just made sure that those were the ones we tested on and made sure that they had a fair distribution of different [outcomes]. … We asked domain specialists, 'Is this an accurate picture, or is it not?' and they said it was, so that's what we [decided we were] going with."*

The Machine Learning Lab's methods for controlling training data enable it to trace changes in an AI application's behavior. This is especially important for countering "data drift"—changes (or drift) in underlying data-generating processes that mean an AI application trained on historical data alone is unable to produce equally valid outputs as the future unfolds. The DBA has experienced some data-drift problems as fraudulent companies change their behavior over the years. For example, the strategies that sham companies use to commit tax fraud tend to evolve over time. This problem needs to be addressed by critical evaluation of training data and possibly by revising or updating the data used.

Responding to the challenge of data (and concept) drift also has implications for the choice of model. Developers can choose from a wide spectrum of models, ranging from offline batch-learning models, which treat data as a static pool and become smarter only when given a new batch of data to learn from, to online self-learning models that learn autonomously from a growing pool of data. For the latter models, the same input can produce different outputs at different times because the system learns "on the fly" and adapts to new information. Online self-learning models can be an appealing option for countering data drift, because of their ability to adapt. "Daniel," a case worker who uses the Company Registration AI application, said that *"we would very much like models that tell us, 'Look at these areas,' areas we didn't even think about. 'Look at these because … there [seems to be] something rotten going on here.'"*

To retain control over training data, the DBA has opted for batch training, not self-learning. This approach to controlling training data helps it to minimize uncertainty stemming from the data and aids in evaluating the outputs from partly or entirely inscrutable systems. In the words of "Jason," a team leader at the Machine Learning Lab. *"… we have made a conscious decision not to use self-learning technologies—i.e., that we'll train a model [on a certain dataset], and then we accept that it will not become smart until we retrain it."*

Controlling Input and Output Data (Dimensions 4 and 5)

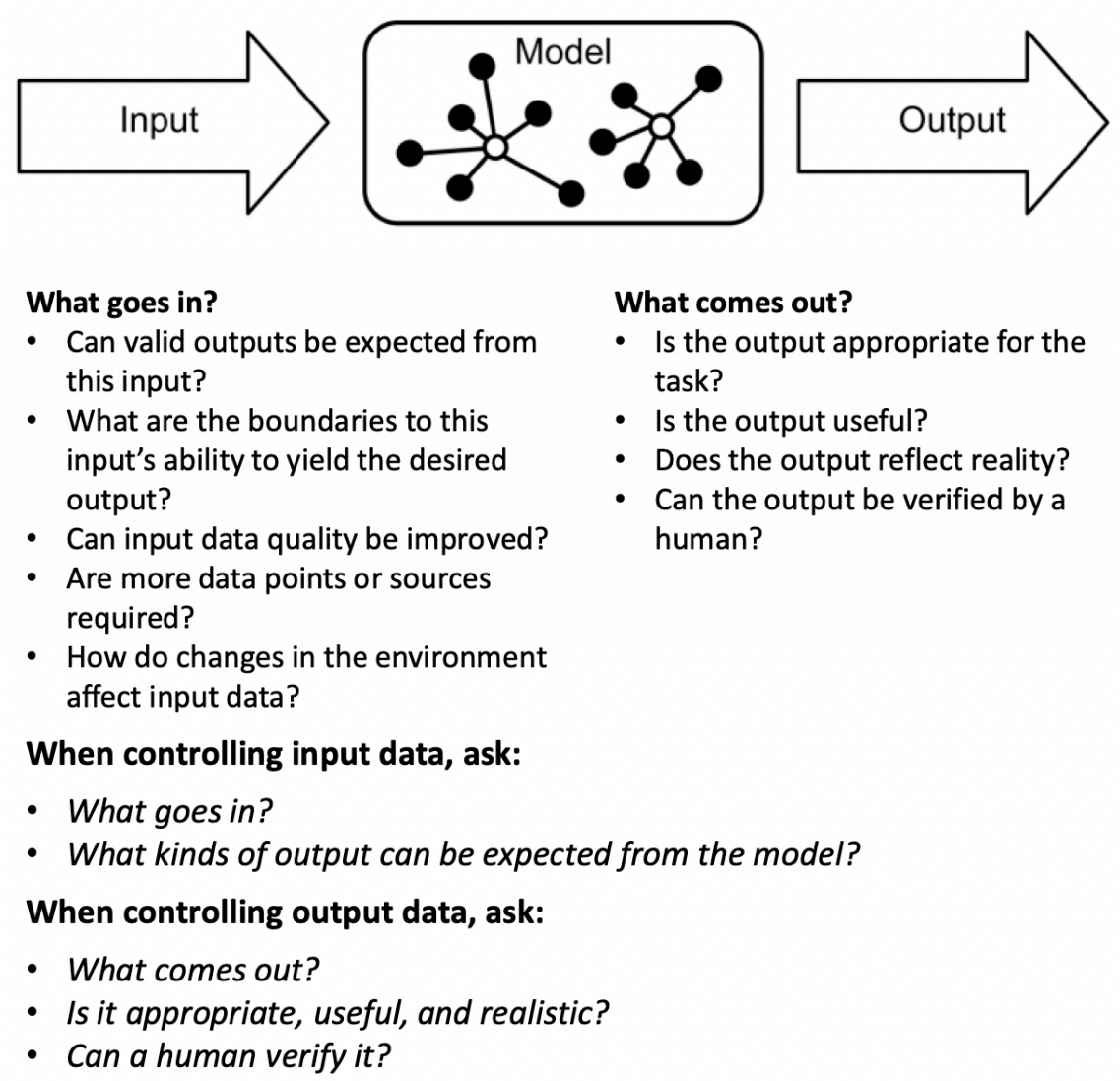
At the DBA, controlling input and output data focuses on understanding what goes into and what comes out of an AI application (see Figure 5). Similar to controlling training data, input control emphasizes the quality of the data processed by the model. In the words of Jason:

> *"When we have a good understanding of where our data comes from, what has influenced [that] data, the causal relation between [input and output data], we understand where, how, and why something happened."*

Low-quality input data can lead to biased or nonusable outputs even if the model has been properly trained. In some cases, the DBA has been able to improve the usability of the output data by preprocessing the input data. For instance, in a project involving citizen-uploaded photos of personal identification documents, rotating the photos before feeding them into the model improved the model's performance significantly.

Controlling output data involves verifying the results produced by an AI application. These

Figure 5: Factors to Consider When Establishing Controls for Input and Output Data

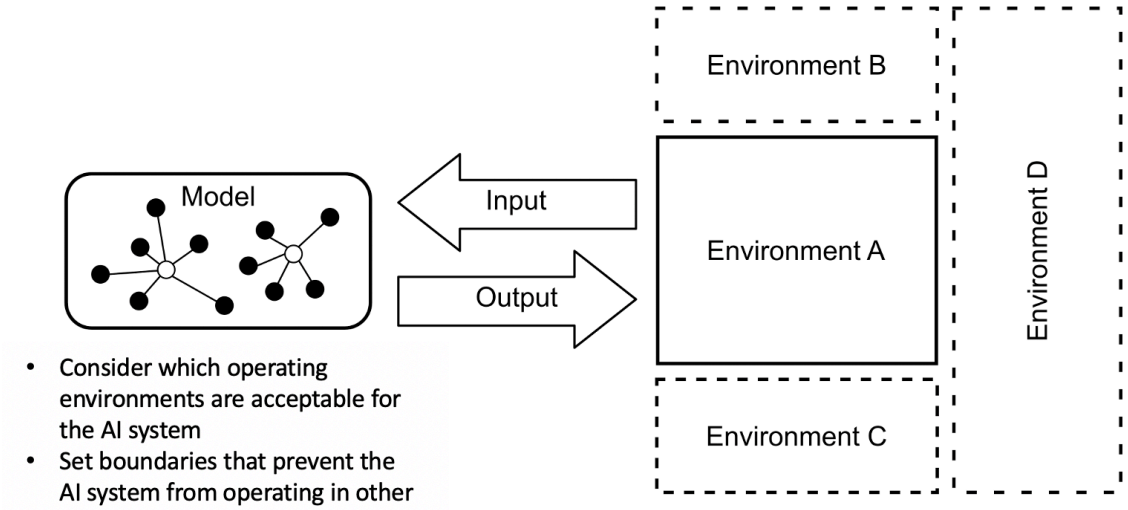


actions may be automated or done manually. For example, case workers using the Signature application manually check documents that the AI application judged to be incomplete, and where the model has expressed low confidence in the correctness of its decision. These outputs are very easy to verify—case workers can determine the completeness of a document by glancing through the relevant fields. This demonstrates that human verification of the output from an AI application does not always require special knowledge of the inner workings of the model, as Steven explained:

> *"If a person calls and asks, 'Why was my document rejected? then a case worker will say, 'That's because you haven't signed it.' 'How do you know that?' 'I have looked at the document. It isn't signed.' So they don't have to answer, 'Well, the neural network said it's because of a variable 644 in the corner.' That's why you can get away with using a neural network in this case, [even though you can't explain how it works]."*

Controlling the inputs to and outputs from an AI application allows the use of inscrutable

Figure 6: Factors to Consider When Setting the Environment Boundaries for an AI Application



models that are hard to explain, provided the organization has the ability to judge the quality of the input and output data.

Setting an AI Application's Environment Boundaries (Dimension 6)

An environment-centered approach to explaining the behavior of an AI application consists of setting clear boundaries for the application's area of operation (see Figure 6). One of the external vendor's testers of the Signature application discovered that the algorithm trained to detect signatures on scanned images of the documents required to found a company would accept an image of a wooden toy animal as valid input and classify it as a signed document. In other words, the application was operating outside its intended environment. To ensure that the application only operated within its appropriate boundary, the DBA created a filter to determine whether the image received is indeed of a document before the image is input into the AI system.

To simplify boundary setting, the DBA designed a software architecture comprising many simple models that operate in highly specific areas, performing very specific actions. This architecture confines each AI application to a limited area, within which its outputs can be easily analyzed. This architecture also limits the damage a malfunctioning application can cause, because the impact is contained in one area. Jason explained, *"By having an event-driven architecture, you can rely on loosely coupled systems, and having sound metadata will help you create order in the chaos of different systems interacting with the same data."*

The architecture also offers a safe and legally compliant way of using black-box AI systems where necessary. The fact that none of the DBA's AI applications make any final decisions affecting citizens or organizations imposes operational boundaries for the applications and also links boundary setting with output control. In many AI applications at the DBA, users have some degree of control over the extent of the operating

environment. For example, in the Company Registration application, case workers are able to adjust critical thresholds for the application to make sure they yield the most useful and precise recommendations possible. Jason explained that this has also facilitated workers' acceptance of the models: *"I was surprised to see the idea of [a] control tower. The ability to mute a model or change the threshold has been a major cultural factor in [the] business adoption of this technology."*

In summary, the environment boundaries of the DBA's AI applications are set through combinations of technological mechanisms (e.g., system design) and managerial controls (e.g., of case workers). However, expert users can gradually develop better rules and tune the boundary thresholds. It is also noteworthy that the boundary for an AI application need not coincide with the boundary between the organization and the external environment. An internal boundary can limit an AI application's effect on the organization's internal operations. For example, the outputs from the Signature project are passed on to another internal agent who continues the processing of the documents deemed by the application to have a valid signature.

In conclusion, the DBA case demonstrates that taking account of all six dimensions of the framework for explaining the behavior of AI systems enables the successful and responsible deployment of various AI applications, even black-box algorithms that are not technically explainable.

Recommendations for Explaining the Behavior of Black-Box AI Systems

Based on our analysis of the DBA case, we provide four recommendations for practitioners. These recommendations encompass both managerial and technological approaches for tackling the challenges of explaining the behavior of black-box algorithms used in AI applications.

1. Implement Strict Controls on the Use of Black-Box AI Systems

Taking account holistically of all six dimensions of the framework described above enables the use of inexplicable black-box AI systems without compromising the safety of operations. Decisions to use such systems depend on the application context and on whether a comprehensive set of control measures is available for the specific application. For instance, Jason stated that a neural network *"has a higher degree of precision but [lacks] transparency; ... we only apply them in areas with low impact or an otherwise objective relation to falseness."*

Our analysis indicates that the use of a black-box AI system, such as a deep neural network, is permissible if:

1. There is minimal possibility of the inscrutability of the system resulting in increased hazards for human stakeholders' wellbeing
2. Using a black-box system does not violate any laws that require the workings of the system to be explained to users
3. The impact of the AI system can be strictly bounded within an internal environment and its output can be controlled by humans.

For instance, the DBA's Signature application rejects a document only if a human can verify the AI application's decision as valid and assume responsibility for the actions that follow. The involvement of human workers can make this approach costly, but the benefits for the DBA, mainly in the form of efficiency gains, have outweighed the additional costs. The DBA case workers can easily screen the problematic documents out of the workflow and devote their cognitive capacity to higher-level activities.

2. Use Modular Design to Make it Easier to Explain the Behavior of an AI System

Breaking complex business processes into smaller modules that can be supported by narrow and well-defined AI applications can make it easier to control and explain the outputs of the applications. For example, designing an AI application to operate a specific function within a process, rather than making it responsible for the entire process, helps to guarantee that it does not—and indeed cannot—obtain data from environments that it should not touch. This means that the developers and users of such AI applications have a high degree of control over

the application's functionality throughout the development and deployment process. They can have greater confidence in the application's outputs and are well placed to detect deviations early on and diagnose any problems that might occur. In essence, modular design of AI systems is akin to a divide-and-conquer approach: rather than try to create an entire explainable system, it is easier to start with multiple explainable pieces that together constitute a bigger AI system. Jason described the DBA's approach as *"feeding the dragon one little biscuit at a time, so we can design models that can be brought into production."*

3. Avoid Online Learning if the Need for Explanation is a Priority

Online learning is appealing for AI applications that have high needs to adapt to environmental changes, but it makes it more difficult to monitor and explain how such an application functions. Online learning therefore results in a reduced level of control and may even prove dangerous in some high-stakes applications. An AI application that learns while operating poses a risk of introducing bias that is not evident from the original design of the system, and that could be challenging to detect and rectify. Difficulty in testing and understanding the behavior of AI systems that use online learning makes it harder to explain how they produce their outputs.

The DBA opted to train its AI applications in a controlled, stepwise manner. This approach protects the applications from the unintended "overfitting"[23] and bias that less controlled learning mechanisms could easily introduce. Note, however, that there is a clear tradeoff between the adaptiveness of the learning mechanism and improved explanation capabilities resulting from offline training. Without subsequent online learning, AI applications trained via offline data may not remain current:

> *"... control departments would rather say, 'We have seen one case that looked like this. Dear machine, find me cases that are exactly the same.' And we have tried to tell them that 'that's fine—we had a case years ago where there were a lot of bakeries that committed a lot of fraud, but now it doesn't make sense to look for bakeries anymore, because now those bakeries are selling flowers or making computers or something different.'"* Daniel, user of the DBA's Company Registration AI application

4. Facilitate Continuous Open Discussion Between Stakeholders

The first three recommendations raise important questions concerned with ethics and responsibility, such as how to determine what is considered biased and who should have the final say in this. We therefore recommend that organizations involve various stakeholders, with distinct perspectives and expertise in the development of AI applications. Beware, though, that involving stakeholders with different backgrounds, approaches and work roles may create obstacles to their ability to communicate with each other. Mark, a data scientist at the DBA Machine Learning Lab, explained:

> *"I think the difficult part has been to get the dialogue with the case workers, who see the world in a different way. ... What exactly is it we should feed the model for getting good predictions, and how do we get the information from the case workers?"*

Communication barriers can be overcome by facilitating further discussion through workshops that involve multiple stakeholders. For example, a data scientist's ability to explain the relevant AI algorithm to domain experts serves as a Litmus test for the ease with which the workings of an AI system can be explained. The DBA's efforts to facilitate dialogue between data scientists and domain experts increased understanding on both sides. The data scientists were able to incorporate important domain-specific factors into the design of AI applications, and the domain experts simultaneously became more informed about the structure of the applications and their operational boundaries.

In addition to focusing on the expected effects on internal stakeholders, the discussions should also consider the implications of using AI systems for the wider business community, economy and society in general. At the DBA, mechanisms such as steering committee reviews

23 Overfitting is where a model accurately describes random errors in the current data to an extent that results in poor fit with future input data.

improve the management of critical ethics-related repercussions that tend to accompany the introduction of AI technologies.

Concluding Comments

There is a compelling need to be able to explain how AI systems operate, and much of the current research on the challenges organizations face in implementing AI is focused on this area. Many recent media reports attest to the disruptive, trust-eroding effects that irresponsible AI implementation can have on organizations and on society. At the same time, advances in AI technologies make it increasingly difficult to develop cutting-edge AI applications whose algorithm-driven decision making can be easily explained.

The Danish Business Authority case study reported in this article provides fresh insights for organizations that want to responsibly deploy complex AI systems in their operations. Some elements of the DBA's approach to making AI systems more explainable are visible in various other organizations, at least tacitly: paying greater attention to the quality of training data and using human oversight to control outputs are now common practices. Our analysis of the DBA's approach shows that taking account of the six dimensions of our framework for explaining the behavior of black-box AI systems can facilitate the successful introduction of AI. Because the DBA is a public-sector organization, it has especially high transparency requirements and has therefore developed tools and management procedures for explaining how its AI applications reach their decisions.

Private-sector businesses may not feel they have as compelling a need to make their AI systems explainable, so—at least at present—they may find less-comprehensive approaches than the DBA's sufficient. Nevertheless, our four recommendations for explaining the behavior of black-box AI systems are equally applicable to public- and private-sector organizations. All organizations, whether public or private, are under mounting pressure to deploy AI-based applications to improve their efficiency and effectiveness while simultaneously demonstrating accountability and responsibility to stakeholders through their ability to explain the algorithm-driven decision making of their AI applications.

In today's business environment, all organizations face constant changes in legislation, norms, codes of ethics, technologies, strategic goals, and the data they generate and use. The controls, choices and boundaries for AI systems are therefore determined by the circumstances that exist when they are set and must be managed if they are to retain their effectiveness over time. To ensure that suitable resources are available for this task, issues relating to explainable AI must be considered when preparing an AI application for production use and throughout its life. The DBA has adopted just such a practice: at set intervals, there is a review of the activities related to each AI application, and the associated costs are factored in from the implementation phase onward. This practice involves collecting feedback from application users and from data scientists on the algorithms' operation, with the functionality being adjusted accordingly.

Organizations should therefore plan to keep their tools and strategies for explaining the workings of their AI systems current through constantly evaluating and retraining their AI systems. We believe the four recommendations we have provided for using the framework for explaining the behavior of black-box AI systems will help organizations effectively address the caveats of such systems while still reaping their significant performance benefits, both now and into the future.

Appendix A: The Danish Business Authority Case Study

Between August 2018 and January 2020, we collected interview and observation data at the DBA. The data was obtained and analyzed through an iterative four-phase process (see the table below), with the phases overlapping and earlier phases informing subsequent ones. We sought to interview a wide range of employees and managers, at several levels in the DBA and with a wide range of tenure, to ensure the data was not biased by the views of long-term or more recent employees.

Phase 1 was largely exploratory and established research collaboration and identified research questions. Phase 2 focused on obtaining in-depth knowledge of the DBA's AI projects and the actors involved. Phase 3 focused specifically

The Four Data-Collection Phases

Phase No. and Data-Collection Theme	Method	Duration (minutes)	Interviewees' Pseudonyms and Roles	Focus of Outcomes
1. Machine Learning Lab Projects Overall	Group interview	105	James (team leader/chief data scientist); Mary (chief consultant, in Annual Reports)	Responsibilities of the DBA; organization structure
2. Machine Learning Lab Functions	Personal interview	90	James	The role of explainability in AI projects; allocation of tasks among stakeholders (the Machine Learning Lab, implementation unit and case workers)
	Group interview	83	David and John (both Early Warning Europe case workers)	
	Personal interview	70	Daniel (an internal case worker in Company Registration)	
	Personal interview	59	Steven (a data scientist)	
	Personal interview	51	Mary	
	Personal interview	116	James	
3. Explainability in AI Projects	Personal interview	51	Steven	Practical means to address explainability issues; the sociotechnical environment of model development
	Personal interview	54	Thomas (a data scientist)	
	Personal interview	50	Linda (a data scientist)	
	Personal interview	48	Michael (a data scientist)	
	Personal interview	52	Mark (a data scientist)	
	Personal interview	53	Joseph (a data scientist)	
	Personal interview	54	Jason (a team leader)	
	Personal interview	48	Susan (a data scientist)	
	Personal interview	62	William (an internal case worker in Company Registration)	
	Personal interview	54	Daniel	
4. Verification of Interpretations from Analysis	Personal interview	55	Jason	Validation of interpretations via interviews and an assessment exercise involving project template mapping
	Assessment exercise	N/A	Steven; Mary; Thomas; Linda; Michael; Mark; Joseph; Jason; Susan	

on explainability and involved all the Machine Learning Lab's employees and two case workers. Finally, Phase 4 focused on validating the interpretations from the analysis of the data collected and gaining fuller insights into the technical infrastructure supporting the lab. Data scientists participated in an assessment exercise with the authors by mapping a descriptive framework for every project conducted by the lab.

The interviews were recorded and then transcribed into 153,195 words of text. The interview data was supplemented with observations carried out by the authors and by document analysis. One of the authors, who has previously worked at the DBA, kept a field diary,

Appendix B: AI Projects at the DBA

Project Name	Project Description
Auditor's Statement	The Auditor's Statement algorithm speeds up verification that the valuations of company assets given in an auditor's statement are correct and that the statement does not include violations. The algorithm is used by internal DBA case workers.
Bankruptcy	The Bankruptcy algorithm predicts company distress and insolvency. It ties in with the Early Warning Europe (EWE) initiative. The algorithm is used by external consultants in the EWE community in Denmark and elsewhere in the European Union. The DBA is not responsible for actions and consequences related to this tool.
Company Registration	The Company Registration algorithm aims to detect fraud-indicative behavior among newly registered Danish companies. The algorithm is used internally by DBA case workers.
Land and Buildings	The Land and Buildings algorithm predicts violations of accounting policies related to property holdings and long term investments. The algorithm is used by internal DBA domain experts.
Passport	The Passport algorithm expedites processing of the documents submitted, supplying a text string from the machine-readable portion of a passport and comparing it with input data from the user. The algorithm is used by internal DBA case workers.
Recommendation	The Recommendation algorithm improves the user experience of the DBA's Virk portal by focusing on personalized content and optimized interfaces. The algorithm improves the portal's usability for external customers.
Sector Code	The Sector Code algorithm speeds up verification of a company's industry-sector code. As of the third quarter of 2020, 25% of company codes were incorrect. The algorithm is used by internal DBA case workers.
Signature	The Signature algorithm, in combination with the associated document filter, speeds up verification of whether company founding documents are signed. The algorithm is used by internal DBA case workers and returns three probabilities: whether the document is physically signed, whether it is digitally signed and whether the signature is missing.

recording observations and taking notes from informal conversations and meetings. This diary dates back to September 2017, when most of the projects were just beginning.

About the Authors

Aleksandre Asatiani

Dr. Aleksandre Asatiani (aleksandre.asatiani@ait.gu.se) is an assistant professor in information systems in the Department of Applied Information Technology, University of Gothenburg, Sweden. He is also affiliated with the Swedish Center for Digital Innovation (SCDI). His research focuses on artificial intelligence, robotic process automation, virtual organizations and IS sourcing. His work has been published in leading IS journals such as *Information Systems Journal* and *Journal of Information Technology*.

Pekka Malo

Pekka Malo (pekka.malo@aalto.fi) is a tenured associate professor of statistics at Aalto University School of Business, Finland. His research has been published in leading journals in operations research, information science and artificial intelligence. Pekka is considered as one of the pioneers in the development of evolutionary optimization algorithms for solving challenging bilevel programming problems. His research interests include business analytics, computational statistics, machine learning, optimization and evolutionary computation, and their applications to marketing, finance and healthcare.

Per Rådberg Nagbøl

Per Rådberg Nagbøl (pena@itu.dk) is a Ph.D. fellow at the IT University of Copenhagen doing a collaborative Ph.D. with the Danish Business

Authority within the field of information systems. He uses action design research to design systems and processes for quality assurance and evaluation of machine learning, focusing on accurate, transparent and responsible use in the public sector from a risk management perspective.

Esko Penttinen

Dr. Esko Penttinen (esko.penttinen@aalto.fi) is a professor of practice in information systems at Aalto University School of Business, Finland. He studies the organizational implementation of artificial intelligence, interplay between humans and machines, and governance issues related to outsourcing and virtual organizing. His main practical expertise lies in the assimilation and economic implications of interorganizational information systems, focusing on application areas such as electronic financial systems, government reporting and electronic invoicing. His research has been published in leading IS journals such as *MIS Quarterly, Information Systems Journal, Journal of Information Technology, International Journal of Electronic Commerce* and *Electronic Markets*.

Tapani Rinta-Kahila

Dr. Tapani Rinta-Kahila (t.rintakahila@uq.edu.au) is a postdoctoral research fellow at the UQ Business School and Australian Institute for Business and Economics, University of Queensland, Australia. His Ph.D. in information systems science was awarded by the Aalto University School of Business. His research addresses issues related to the decommissioning of IT systems, organizational implementation of artificial intelligence and automation, and the dark side of IS.

Antti Salovaara

Dr. Antti Salovaara (antti.salovaara@aalto.fi) is a senior lecturer at Aalto University, Department of Design, Finland, and an adjunct professor in the Department of Computer Science, University of Helsinki. He studies human-AI collaboration and online trolling, and the methodology of user studies. His research has been published in conference proceedings such as CHI, and in leading journals, including *Human Computer Interaction, International Journal of Human-Computer Studies, MIS Quarterly* and *European Journal of Information Systems*.