SAP THOUGHT LEADERSHIP PAPER | PUBLIC
Public sector

# Building Explainability into Public Sector Artificial Intelligence

## Helping Stakeholders Understand the Thinking Behind AI Decision-Making

THE BEST RUN SAP

# Table of Contents

**All over the world, every day, public sector organisations are adopting artificial intelligence (AI) for both internal and citizen-facing government services. A crucial consideration with building AI systems relates to human decision-makers' ability to understand and explain how these systems generate their decisions. This is often called AI explainability.**

**Requirements such as these are codified in the European Union's General Data Protection Regulation (GDPR) that reserves the right for individuals to access meaningful explanation on decisions that affect their lives.**

# The Context

The emergence of complex, advanced algorithms is making it increasingly difficult to explain the AI models' inner workings[1]. For instance, deep-learning systems learn autonomously from data and propagate their learning across many layers of a network. This renders it impossible for even seasoned data scientists to trace the rationale behind the algorithmic decisions. That's why a recent report on AI challenges[2] concludes: 'For public sector organizations that have to practice high levels of transparency and accountability, this lack of explainability represents a significant roadblock for AI implementations.'

Current research on this topic demonstrates that explainability extends beyond technical traceability for AI models, to meaning. This requires the consideration of several types of explanation, aimed at various stakeholders, who differ in their reasons for using the system, or being subject to its decisions and actions[3]. Even when the internals must remain partly inscrutable, a close look at the training data, input, and system boundaries can greatly improve explainability[4].

While the need for explainability is clear, awareness of how organisations should go about facilitating it is still in its infancy.

[1] 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 (2021), 325–352.

[2] Rinta-Kahila, T., Someh, I., Indulska, M., et al. Delivering AI Programs in the Public Sector: Guidelines for Government Leaders. The University of Queensland and SAP SE, (2020).

[3] Ribera, M. and Lapedriza, A. Can we do better explanations? A proposal of user-centered explainable AI. CEUR Workshop Proceedings, 2327, (2019).

[4] 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 (2020), 259–278.

# A Fresh Perspective

In this thought leadership paper, we will offer a fresh perspective for managing AI explainability in the public sector. Over the following pages, we present explanations as an approach for engaging with diverse stakeholders while developing and implementing AI.

This process, aimed at aligning AI operations with stakeholder-specific perspectives and knowledge, encourages multiple iterations and feedback loops in which AI models' learning is compared with the knowledge possessed by domain experts, and the actual reality.

Through the process of 'explaining' AI, public sector organisations can achieve three main outcomes:

- Domain experts will gain **new technical skills**, enabling them to work with the AI-informed systems more effectively, inhibit knowledge loss, and develop professionally
- AI systems will **grow increasingly accurate**, representative, and meaningful through iterative and continuous stakeholder engagement and critical comparison of different AI models
- Public-sector **processes and services will benefit** from AI-based systems and an AI augmented workforce, which, in turn, will enable governments to better meet the needs of citizens and employees alike

# Our Research

To gain deeper understanding of how AI systems can be developed in a way that cultivates explainability, we studied two AI projects undertaken in Australian public agencies. The systems, which we will refer to as **Tax AI** and **Health AI**, were developed to help improve public-facing services by exploiting the prediction capabilities of machine-learning models:

The **Tax AI** project was aimed at identifying taxpayers who showed a risk of becoming chronic debtors by failing to make their payments on time. The state tax revenue-management office turned to AI technology in its search for ways to 'improve taxpayer services and overall (tax) debt collection rates'. They had a vision of AI giving the office's call-centre workers access to richer and more accurate insights so that they could implement appropriate intervention strategies tailored to individual taxpayers' circumstances.

A proof of concept was developed for land-tax debt, on account of the 'high rates of payment default experienced in this area.' The model was trained with roughly 200 million data records, of nearly 100,000 taxpayers, spanning seven years. It correctly predicts over 80% cases of taxpayers entering debt.

**Health AI** was aimed at identifying patients at risk of developing sepsis while waiting for treatment in a hospital's emergency department (ED). Many patients arriving at the ED are susceptible to this life-threatening condition, which is caused by the human body responding to infection in a way that damages its own tissues and organs.

While hospital personnel can treat sepsis effectively at low cost with antivirals and antibiotics, detecting it in a patient early enough is far from straightforward: a combination of various confounding factors and humans' cognitive limits lead to unnecessary deaths from the condition. Seeking its timely identification and treatment, a state health department developed an AI system to detect signs of sepsis in ED waiting-room patients and alert triage nurses to the possible need for rapid treatment.

For both projects, our in-depth case studies used data from interviews with key stakeholders in the AI development and from written documentation. The conceptual underpinnings of our interpretation of the data are presented next.

# Managing AI Explainability

At the heart of any AI technology is an AI model, an algorithm trained with data that mimics human decision-making processes. The models are an abstract representation of some portion of reality and are designed to predict domain-specific realities (for example, patients at risk of sepsis). With them come three inevitable gaps in understanding and performance (see Figure 1), all of which need be addressed and managed[5].

**Gap 1** – Inconsistency between the user's understanding and the AI model:
- Users do not fully comprehend the AI model's logic
- The gap inhibits user trust and stifles acceptance of AI

**Gap 2** – Inconsistency between the AI model and reality:
- The AI model is not a complete or comprehensive representation of reality
- The gap causes poor model performance or produces bias against specific cohorts

**Gap 3** – Inconsistency between the user's understanding and reality:
- Users exhibit bias or lack of understanding related to how things work in reality
- The gap reduces domain experts' potential real-world impact

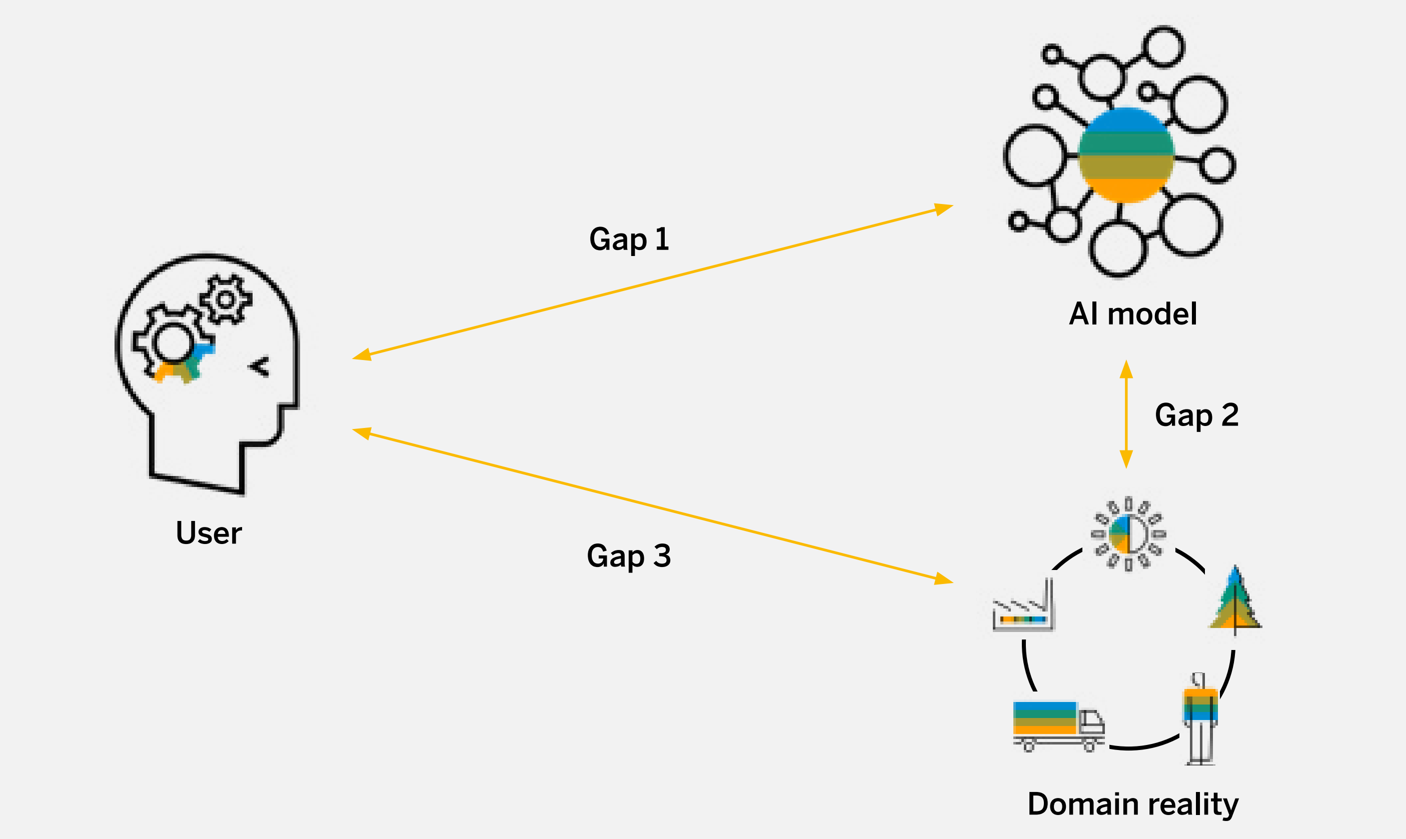

Figure 1: The Gaps Between the User, AI Model, and Domain Reality

[5] Kayande, U., De Bruyn, A., Lilien, G.L., Rangaswamy, A., and van Bruggen, G.H. How incorporating feedback mechanisms in a DSS affects DSS evaluations. Information Systems Research, 20, 4 (2009), 527–546.

AI explanations can help bridge these gaps by aligning the AI model with reality and the minds of its users. We identified three pathways by which this occurs, outlined in Figure 2, showing that explanations:

1. Aid in upskilling domain experts
2. Enhance the AI system's performance
3. Improve processes and services

The case studies led us to an understanding of the role of explanations for each of these, and we developed a sense of how explanations enabled these outcomes in both projects.

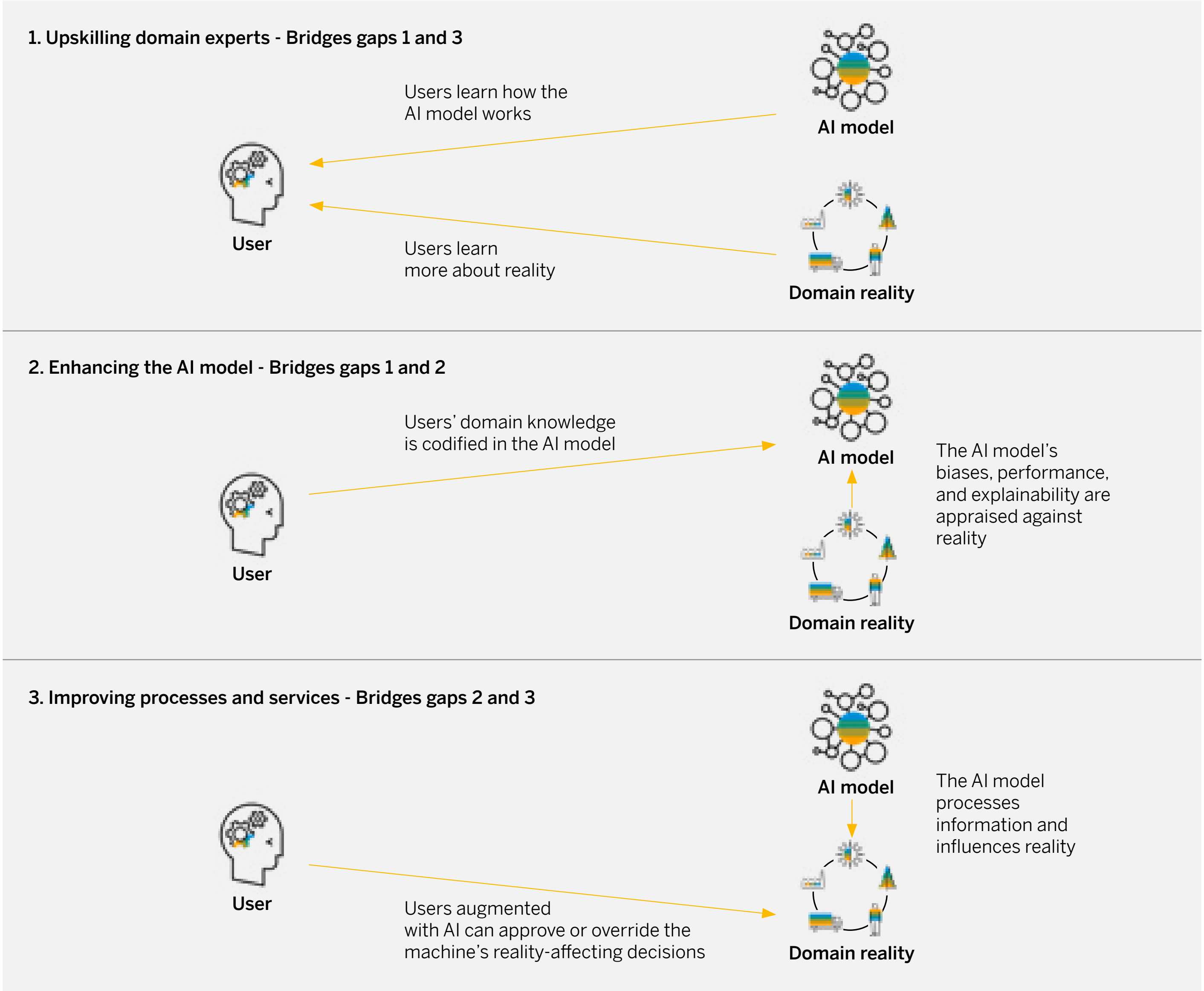
1. Upskilling domain experts - Bridges gaps 1 and 3
Users learn how the AI model works
AI model
User
Users learn more about reality
Domain reality
2. Enhancing the AI model - Bridges gaps 1 and 2
Users' domain knowledge is codified in the AI model
AI model
The AI model's biases, performance, and explainability are appraised against reality
User
Domain reality
3. Improving processes and services - Bridges gaps 2 and 3
AI model
The AI model processes information and influences reality
User
Users augmented with AI can approve or override the machine's reality-affecting decisions
Domain reality

Figure 2: The Role of Explanations in Bridging the Gaps

## USING EXPLANATIONS TO UPSKILL DOMAIN EXPERTS

Explanations serve as a medium of education and professional development for an AI model's domain-expert users. Typically, domain experts are neither conversant with data driven decision-making nor aware of how machine-learning algorithms work.

In general, domain experts lack certain technical skills relevant for understanding and collaborating with AI technology. Their conception of reality may be skewed too, due to their subjective experience and biases. As technology is increasingly carrying out daily tasks, domain experts may also experience deskilling by forgetting some of the factors that contribute to the process[6].

In both projects, explanations provided domain experts with upskilling opportunities by:

1. Developing their technical skills by learning how AI models learn and produce results
2. Refreshing and deepening their domain expertise by providing evidence-based insights into how domain areas operate

This process also aided in managing employee fears related to AI technology and facilitated the systems' acceptance.

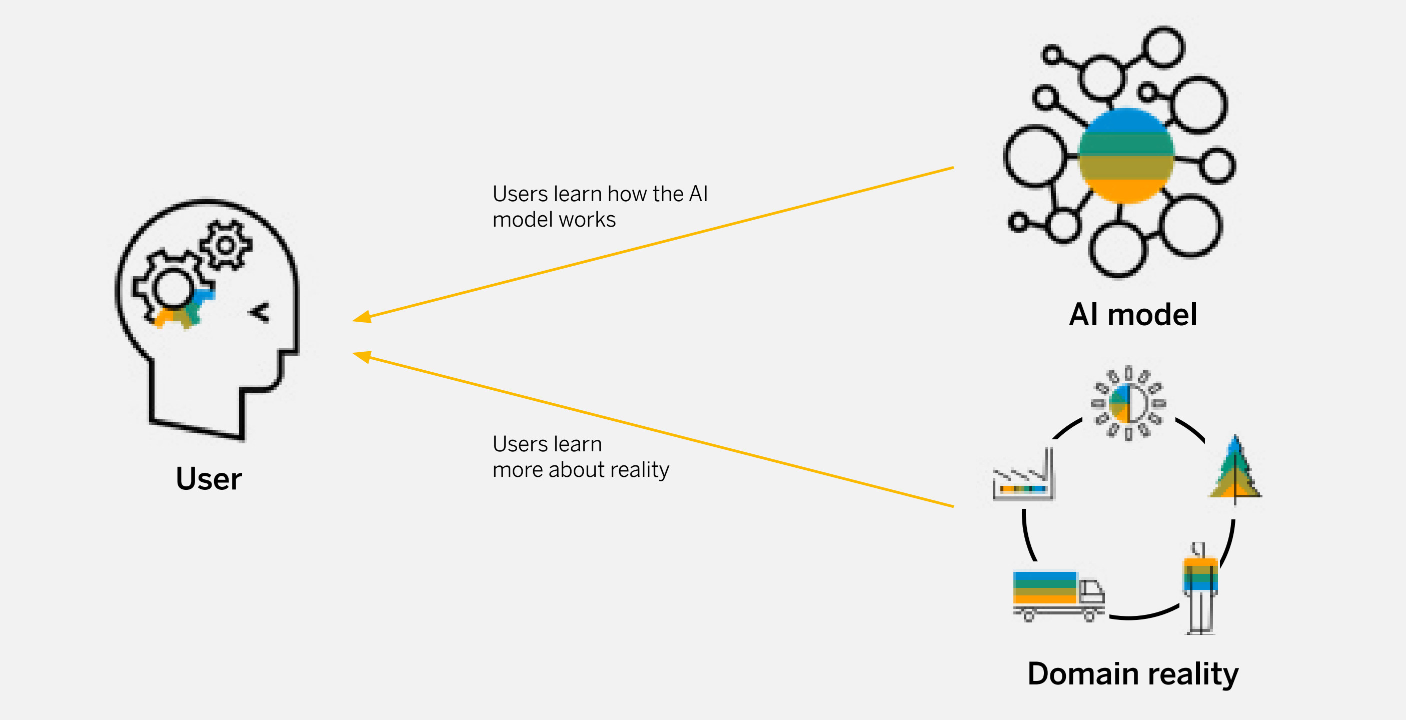
Users learn how the AI model works
AI model
User
Users learn more about reality
Domain reality

Figure 3: Using Explanations to Upskill Domain Experts

[6] Rinta-Kahila, T., Penttinen, E., Salovaara, A., and Soliman, W. Consequences of Discontinuing Knowledge Work Automation – Surfacing of Deskilling Effects and Methods of Recovery. In Proceedings of the 51st Hawaii International Conference on System Sciences. 2018, pp. 5244–5253.

In the Tax AI project, data scientists began educating tax experts on the principles that formed the foundation of the AI system's design including its input data, functionality, limitations, and output decisions. The team had to explain the overall logic of the AI's operation so that customer-service staff could understand how the model takes various factors into consideration and recognise the probabilistic nature of its outputs. Illustrations of specific scenarios, with visualisations, enriched the team's description of how the AI reached decisions about taxpayers.

By 'unwrapping' the AI decision-making process, the data scientists were able to deepen the domain experts' understanding of why and when taxpayers defaulted on their debts.

For example, people who have been on interest-free payment plans in previous years tend to become debtors later by failing to pay by the due date without any apparent reason. The AI system's customer-journey visualisations helped the staff understand that many of these people actually believed that they were still on extended payment plans and thus thought they were acting in accordance with expectations.

Demonstrating the AI's logic also highlighted to tax officers the need to collect more detailed textual data from customer interactions so that the model could be still more sensitive to the reasons for taxpayers' delayed payment. Tax AI's business architect explained that the customer-service staff 'had to understand what they were doing. How do we now capture data, how do we now use this tool?'.

A similar technique was employed in the Health AI project.

Hospital nurses had been relying on simple cut-off values to identify whether a patient was likely to have sepsis. As the clinical director explained: 'people would look at patients and (declare that) a pulse of more than 120 is sepsis-positive (so) pulse 119 is not sepsis-positive.'

These thresholds were problematic because the reality is more complex than binomial cut-offs and reliance on such a simplification renders a nurse likely to miss many sepsis cases. Data scientists and the clinical director explained how AI can help address this complexity. This helped the nurses move beyond the old paradigms and gain more comprehensive understanding of the decision features involved and of their importance for detection of sepsis.

Educating the domain experts on the meaning of data and how the relevant AI models function and utilise the data contributed to a clear improvement in their technical skills (bridging gap 1). Furthermore, as the data-science team took advantage of data and AI's power for exposing and explaining previously overlooked features, these experts' understanding of their domain refreshed and deepened (bridging gap 3).

## USING EXPLANATIONS TO ENHANCE THE AI MODEL'S PERFORMANCE

By explaining the decisions an AI model has made, and the reasons underlying those decisions, stakeholders can compare and contrast model outputs with their domain knowledge and experience. This process of exposing and scrutinizing an AI model's behaviour through relevant and consumable explanations can reveal potential biases or errors and trigger feedback cycles that ultimately enhance system performance and rectify biases[7].

For this to work, explanations need to be tailored for their audience, as different stakeholders (for example, domain experts, managers, citizens) have varying levels of understanding AI and application domains.

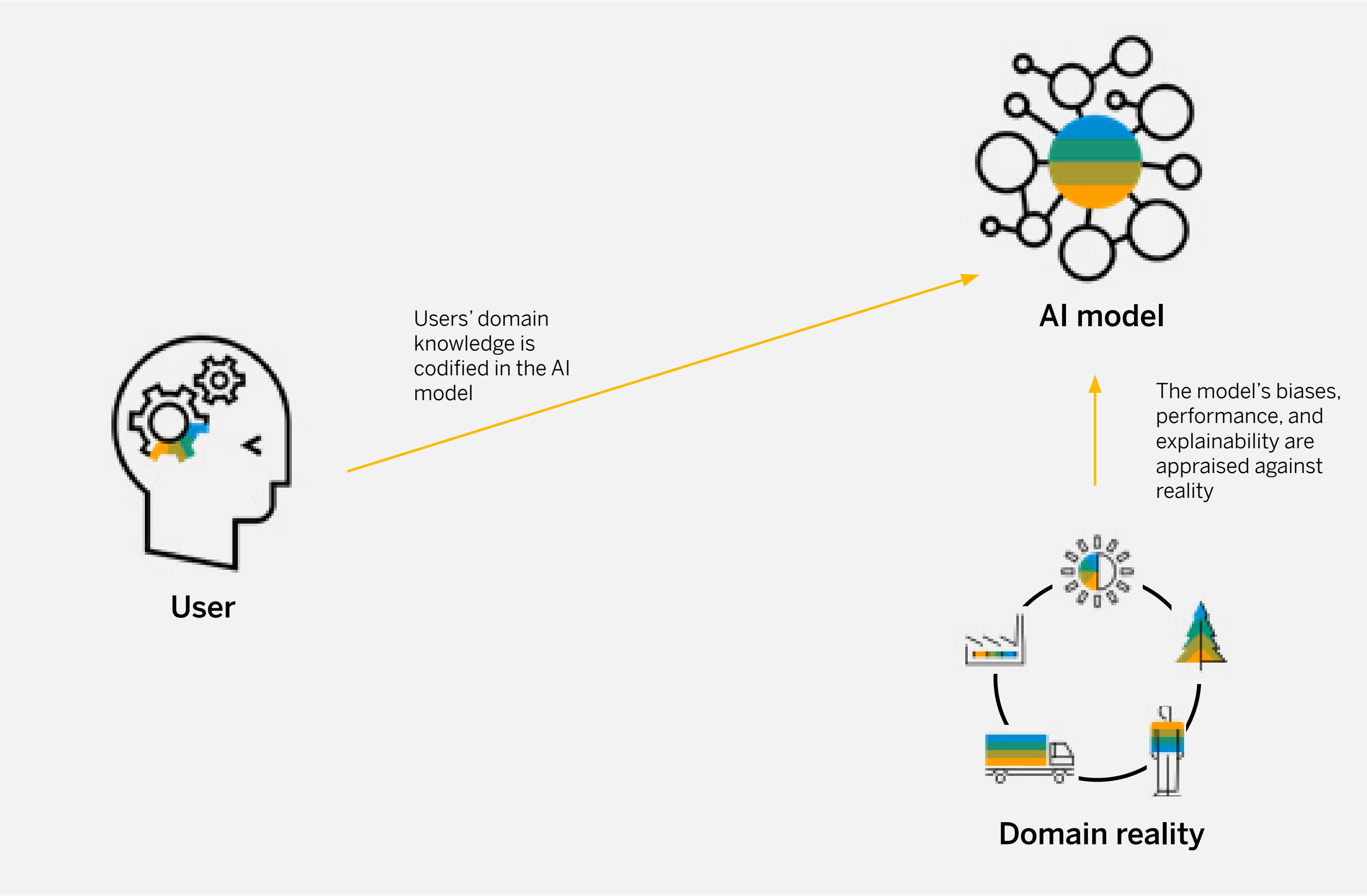
AI model
Users' domain knowledge is codified in the AI model
User
The model's biases, performance, and explainability are appraised against reality
Domain reality

Figure 4: Using Explanations to Enhance AI Models' Performance

[7] Wixom, B., Someh, I., Zutavern, A., and Beath, C. Explanation: A New Enterprise Data Monetization Capability for AI. MIT Center for Information Systems Research (CISR), Working Paper, NO. 443, (2020).

In both projects, explanations played a key role in building and training the AI models and in improving their performance. Since the domain experts were the ones with insight into which factors are signals of a citizen failing to pay tax or a patient developing sepsis, the data scientists worked alongside them to gauge models' outputs against human expertise.

As the data-science team explained how the model worked, domain experts contributed their knowledge to training of the model. This included informing data scientists of different data points, decision variables, or how and why the variables are related to one another, which they could then codify into the AI model, making the models more meaningful.

The interaction between the domain expert and data scientists was mediated by a business architect who drew together the two fields' knowledge. This facilitated explanations that create connections between technical and non-technical stakeholders.

The business architect for Tax AI asked:
*"What does it mean if a taxpayer has three different late payments? Is that a high risk? Is that not? For them to be able to train the model, we need to get all that business context and basically take that from those business users (and) put them into (the heads of) our data scientists. That was sort of from the perspective of us training the model and training the machine."*

Still, as humans, domain experts are fallible and susceptible to bias. Therefore, there was a need to contrast the AI models against reality to make sure they do not exhibit biases caused by human judgement (domain experts' views) or skewed datasets.

When evaluating the models' performance against real-world outcomes, both project teams faced what is known as the explainability / accuracy trade-off: as they moved from simple machine-learning algorithms to more complex ones, such as deep neural networks, the teams noticed that increasing accuracy came with the cost of decreasing explainability.

The Tax AI data scientist said:
*"Even as a data scientist, when we run a neural network, we still don't really fully understand (what is) happening inside the neural network; (...) random forest is much easier, because we can visualise the decision tree and show how the decision is made."*

Because different datasets require different kinds of modelling approaches, the Tax AI team adopted an approach of combining a random-forest and a deep-learning model. This combination of models not only helped boost performance but also enabled them to increase explainability, as the resulting ensemble model was far more interpretable than the largely inscrutable approach of a deep neural network.

The Tax AI data scientist remarked:
*"We tested (...) ensemble methods, and we've just (...combined) some of the algorithms together(...) we ended up using random forest and also the neural network(...) (We) mash the two results together: (for the) neural network, it is hard to explain, but for random forest, we can actually show them (that) these are the feature importance – why the model says it this way."*

In the case of Health AI, the team was highly conscious of the explainability-accuracy trade-off when they explored competing models. The team started with logistic regression because it was *"(the) most transparent and (integrable) with clinical work flows"*, according to the data-analytics director.

As they moved to more complex models, such as boosting techniques and neural networks, they scrutinized both accuracy and explainability with each model. If an inscrutable model seemed to deliver notable performance benefits over an explainable one, the team would try 'to build some transparency back in' by seeing whether linear approximations could be run on top of the inscrutable model.

In summary, as data scientists explained the AI model's inner workings to domain experts, the domain experts learned what key domain knowledge had to be encoded into the model. This encoding process resulted in models that were more relevant and useful to the end users (bridging gap 1). Contrasting the models' outputs against domain reality helped to reveal 'what AI knows'. This, in turn, informed the pursuit of 'good enough' performance and redressing of undesired biases (bridging gap 2).

[8] Rinta-Kahila, T., Someh, I., Indulska, M., et al. Delivering AI Programs in the Public Sector: Guidelines for Government Leaders. The University of Queensland and SAP SE, (2020).

## USING EXPLANATIONS TO IMPROVE PROCESSES AND SERVICES

AI explanations inform real-world decision and policy making by delineating how model outputs will influence processes and services in ways that generate positive outcomes for different stakeholders[8]. This requires making AI's outputs accessible, understandable to the end users, and explicit in how they translate into action and concrete effects.

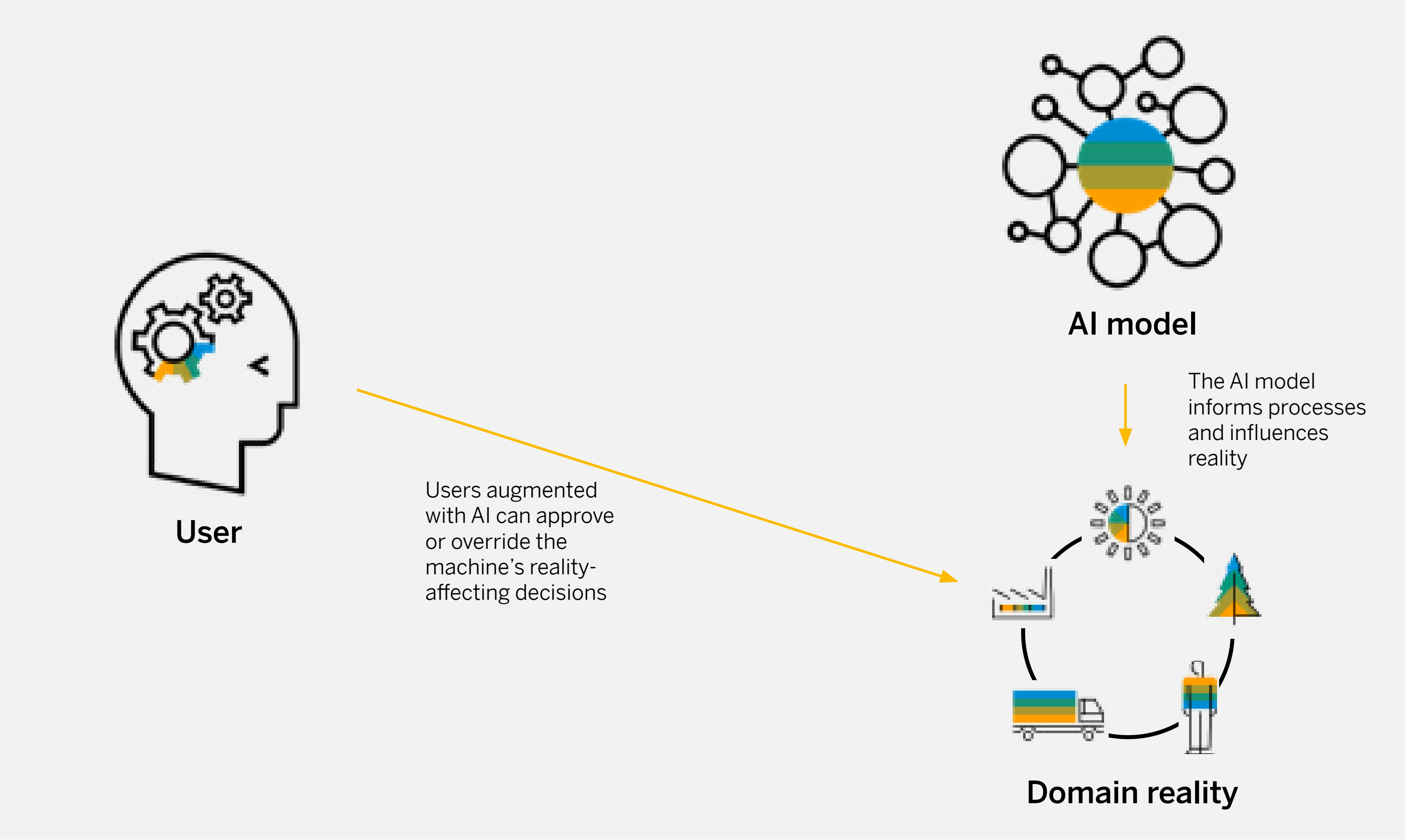
AI model
The AI model informs processes and influences reality
User
Users augmented with AI can approve or override the machine's reality-affecting decisions
Domain reality

Figure 5: Using Explanations to Improve Processes and Services

To ensure that the AI implementations would engender real-world influences, the machine learning (ML) models had to be integrated with existing workflows. Both project teams created simple, user-friendly AI interfaces that present the model's insights and the recommended actions, alongside explanations of their basis.

In the Tax AI project, the data-science team collaborated with an IT application team to develop a user interface that translates the AI's insights into graphical depiction of a customer journey. The interface applies simple textual and visual cues pointing to whether and why a given taxpayer seems likely to become a debtor and, backed up by these explanations, suggestions for actions that may constitute appropriate intervention.

The business architect for Tax AI noted:
*"Like a traffic light, green to red, to show increasing risk to low risks. So really easy to understand, and we just used the percentage: were they 85% or…?"*

Along similar lines, the Health AI interface displayed patients' estimated risk of sepsis by means of distinct colours, alongside explanations. According to the data-analytics director for Health AI:
*"[S]o we had red and orange as sort of different levels in the mock-up, but we're probably going to try just red highlighting to keep it simple for the initial launch of this tool, and we're going to get their feedback on whether there's any value in having a nuanced approach (…) you can click the patient and get the full range of factors that are leading to that prediction."*

Thus, the AI models, by informing decisions and processes that affect citizens, suggest ways to make a real-world impact. Moreover, arming upskilled domain experts with AI interfaces enables them to exert effects in an informed manner within their domains. As the data scientist for Tax AI notes:
*"It basically augments your job. So it helps you in your daily job, to help you (gain a better) understanding about your clients – about your taxpayers, for example. So the next time, when the taxpayers call, you'll be able to understand what they have done in the past, what action (…) had been taken so that you will be able to advise better."*

Explaining the AI model's limitations (for example, probabilistic nature of outputs) to end-users helped to meaningfully keep humans in the decision-making loop, further highlighting the need for establishing human accountability when implementing AI. The business architect for Tax AI elaborated on how this was explained to the users:
*"We told them specifically: You cannot automate everything end to end. For example, when the machine tells 'this is a 90% probability [someone] will become a debtor…', you can't have that without human input and automatically let the machine send letters to the taxpayers. Because the machine itself is not 100% bulletproof."*

In addition to providing more effective responses to domain problems, arming domain experts with AI models can produce more proactive solutions in the long run. As AI explanations help to reveal factors that contribute to undesired outcomes (for example, becoming a debtor or developing sepsis), domain experts can drive real-world changes to products, services, and processes targeted at eliminating those factors.

In Tax AI, some individual customers aged under 35 were failing to pay even after the agency had sent them several letters. This data coupled with recorded absence of digital contact led the agency to change the contact method for this cohort from whitemail to email, which resulted in an increased rate of timely payment.

In both projects, the AI systems' explanatory interfaces guaranteed the models' ability to yield real-world benefits (bridging gap 2). Injecting these AI systems into work processes while building technical competence into domain experts kept them highly involved in the decision-making loop, ultimately helping them to provide better services without ceding control to the AI (bridging gap 3).

# Guidelines for Managing AI Explainability

Our research enabled identifying practices that have potential to facilitate AI implementation in the public sector and so help organisations carry out impactful AI projects with success. Our insights include:

**Build explainability into complex AI models by examining and incorporating alternative traceable models.** While many advanced AI models are inscrutable black boxes, they can be examined and tightly correlated relative to traceable ones. Combining different models may add some level of transparency to the decision-making process and even enhance performance. Following this approach requires heavy scrutiny from the explainability perspective: consistent, reliable decision-making necessitates assessing, comparing, and calibrating the models' performance against one another.

**Move beyond technical traceability to explanations that engage and involve stakeholders in AI-model development.** Our study highlights the importance of considering AI models' users, their knowledge, values, and perspectives when building AI. Training good AI models requires input from stakeholders ranging from domain experts and executives to citizens, for many of whom the concept and potential value of AI remains unclear. However, presentations of model's technical operations and trace will not be meaningful for these stakeholders. Shifting from technical traceability to accessible explanations of the decisions, actions, and mechanics relevant to the stakeholders, can encourage their deep engagement with the model, allowing them to inform the models' training, and guide its evolution. This enables user upskilling and helps to overcome user resistance that typically plagues IT deployments. Requesting stakeholder feedback and incorporating it into the AI model should be done on a continuous basis as part of the overall governance mechanism – not just in the development and implementation phases.

**Integrate AI into work by means of user-friendly explanatory interfaces.** When a model's complexity is mirrored by a highly technical application interface, the system is not a good tool for many non-technical stakeholders. Regardless of how advanced the code behind the interface is, domain experts need simple tools built with specific user requirements in mind: interfaces that deliver an end-to-end process or service experience, preferably with clear explanatory visualisations. The AI models' integration into existing work flows, products, and services proves just as vital. Clear, uncluttered interfaces with visual aids get the most from humans in AI-augmented decision-making processes.

**Educate and empower frontline staff to exploit AI but also override its decisions.** AI models are never quite in complete harmony with the real world, just as any other model deviates from reality. Errors and inaccuracies naturally follow. An AI model is particularly failure-prone when dealing with novel cases or with contexts perhaps not represented in the training dataset. Therefore, humans must remain in the decision loop after the AI system is deployed. Users who are kept educated and informed with explanations that clarify the models' boundaries and limitations add substantial value when empowered to exercise decision-making authority and override AI decisions. Furthermore, models can learn from cases of users questioning or overriding their decisions.

**Plan for an iterative process.** AI technologies are still nascent and emerging. Therefore, their implications for human stakeholders, work processes, and organisational arrangements remain poorly understood. This demands awareness and flexibility: organisations must exercise prudence via an iterative process wherein the business goals behind the AI system are refined and refocused and in which the AI models get scrutinised, both periodically and in response to stakeholder feedback. Explanations are crucial for detection of issues that necessitate revisions to the organisation's AI systems.

# Looking Ahead

The two case studies, Tax AI and Health AI, highlight the need for organisations to embark on an overall learning journey and to think beyond traditional IT projects when implementing AI.

First, AI is changing the nature of work and how it is performed. Businesses that choose to invest in AI will also need to allocate resources for training the workforce of tomorrow. As noted above, managing AI explainability goes far beyond scrutinizing the technical traceability of AI models. While these models gain technical improvements via ML, explainability entails an overall learning experience where both humans and machines accumulate knowledge jointly and iteratively. As organisations and their stakeholders continue to learn about the workings of ever-evolving AI models, they also gain new insights about their own work processes, employees, and customers. This helps them to focus resources for training employees and creating a culture of learning.

Second, traditional functional forms of work organization will not be enough to support developing AI-powered organizations. Our research points toward shortcomings of treating AI projects as traditional IT projects run mainly by the IT department. Due to the less deterministic nature of the AI technology, managing AI projects and the models' explainability requires higher extent of organisational flexibility than traditional IT implementations do. The conventional division between business and IT must be reconsidered, as AI systems require a strong business focus. Being too technically driven risks a one-sided pursuit of performance at the expense of explainability. Different stakeholders should be onboarded to continuously sense and address the AI systems' implications. All this requires developing an organisational capability to learn with and from AI models.

# Report Authorship

**The University of Queensland Researchers**

*Australian Institute for Business and Economics*
Dr Tapani Rinta-Kahila

*UQ Business School*
Dr Ida Someh, Professor Marta Indulska

**SAP advisors**
Ian Ryan, Ryan van Leent

**Acknowledgements**
We are grateful for the anonymous case organizations for their collaboration in data collection.

**Citation:**
Rinta-Kahila, T., Someh, I., Indulska, M., Ryan, I., and Van Leent, R., (2021). "Building Explainability into Public-Sector Artificial Intelligence," The University of Queensland and SAP SE. Link to report: XXXX, DOI: XXXX

Follow us

www.sap.com/contactsap

75691enUS (21/05)

THE BEST RUN SAP®