



ASIA PACIFIC | JULY 2022

AI Transparency & Explainability

A Policy Prototyping Experiment



NORBERTO NUNO
GOMES DE ANDRADE



About Open Loop

Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies.

The program, supported by Meta (previously Facebook), builds on the collaboration and contributions of a consortium composed of regulators, governments, tech businesses, academics and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rulemaking processes in the field of tech policy.

This report presents the findings and recommendations of the Open Loop's policy prototyping program on AI Transparency and Explainability, which was rolled out in the Asia-Pacific region from April 2020 to March 2021, and in partnership with Singapore's Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC).

This work is licensed under a [Creative Commons Attribution 4.0 International License](#).



Cite this report

Andrade, Norberto Nuno Gomes de. "AI Transparency and Explainability - A Policy Prototyping Experiment" (2022), at https://openloop.org/wp-content/uploads/2022/07/AI_Transparency_&_Explainability_A_Policy_Protoyping_Experiment.pdf

Acknowledge- ments

This policy prototyping program was designed and facilitated by **Meta**, and rolled out in collaboration with **Singapore's IMDA** and **PDPC**. We want to thank in particular Yeong Zee Kin (Assistant Chief Executive – Data Innovation and Protection, Singapore IMDA), and Lee Wan Sie (Director – Development of Data-Driven Technologies, IMDA) for their support and invaluable contribution to this project. A very special thanks to Verena Kontschieder (AI Policy Program Manager, Meta) for ensuring the coordination and management of this project, and to Melanie Glöckler for the research assistance provided.

We would like to thank the following companies for their partnership and participation. Without their commitment and active involvement this project would not have been possible:

Bukalapak Indonesia
Deloitte Singapore
Evercomm Singapore
Halosis Singapore
Meta Global
Ngee Ann Polytechnic Singapore

Nodeflux Indonesia
Trabble Singapore
Travelflan Hong Kong
Traveloka Indonesia
Qiscus Singapore
QSearch Taiwan



Acknowledgements

Thank you in particular to the individual experts that represented the participating companies throughout the program:

Even Alex Chandra, Muhammad Ghifary, Muhammad Haris, Raden Brahmastro Kresnaraman, Chun Ping Lim, Ghuizen Chen, Ted Chen, Bharath Mathivanan, Ting Yan, Carsten Hansen, Ramya Sethuraman, Kara Davison, Andrew Hendrawan, Sonja Johar, Adi Kurniadi, Ibrahim Abbas, Kristiyanto, Nadhira Fasya Ghasani, Nelson Liu, Jia Chiun Wong, Evan Purnama, Pitula C. Hariy, Muhammad Md Rahim, Roger Do, Angela Ho Li Ting, Ian Low, Jian Liang, Michelle Teo Yin Zhi, Alex Chen, Wenbin Wang, Doan Lingga, Eric Zhang, Leialoha Lubis, Brinanda Lidwina

Thank you also to our technical partners and event organizers for their support:

- **AI Singapore**, in particular Laurence Liew, Koo Sengmeng, Kevin Oh, Weng Jianshu, Lim Tern Poh, William Tjhi, Najib Ninaba and Maurice Manning.
- **Aicadium**, in particular Feng-Yuan Liu, Cassandra Locke, April Chin, Hien Linh Nguyen, Qiao Han, Kok Meng Tan.
- **Plug and Play**, in particular Carinna Goh and Inhee Choi.
- **Craig Walker Design and Research**, in particular Kernow Craig.
- **TTC Labs**, in particular John Lynch.

Final thanks to all other participating company delegates who have attended or supported their teams in this Open Loop program.



Contents

Foreword	6
----------	---

Executive summary	8
-------------------	---

Introduction: AI Transparency and Explainability	15
--	----

Understanding the concepts and their relevance	16
--	----

Increasing societal relevance and technical advancements	17
--	----

Explainability as a vehicle for Responsible AI and as a target of regulatory focus	17
---	----

The Singapore AI Explainability policy prototyping program	19
--	----

Project overview	20
------------------------	----

Project goals	21
---------------------	----

Partners	22
----------------	----

Participating companies	23
-------------------------------	----

In Conversation with AI Singapore	27
---	----

Methodology	28
-------------------	----

Foundational Phase:	
Evaluation and Testing of the Policy Prototype	34

Policy clarity	35
----------------------	----

Policy effectiveness	36
----------------------------	----

Policy actionability	39
----------------------------	----

Procedural Phase:	
The technicalities and the trade-offs of building AI explainability solutions	40

Technical assistance provided to the participants	41
---	----

In conversation with Aicadium	42
-------------------------------------	----

Trade-offs involved in building Transparency & Explainability	43
---	----



Contents

Delivery Phase:	
Presenting and Communicating the AI Explainability solution	45
XAI-related objectives	46
Technical Considerations	46
Policy considerations.....	47
Usability considerations	49
Challenges from this exercise, learnings for future ones.....	54
Policy Insights and Recommendations	56
"Get practical"	57
"Get personal"	58
"Connect the Dots"	58
"Get creative together"	59
"Test and Experiment"	60
Conclusion	62
Endnotes	64
Bibliography	68
Annexes	71
AI Transparency and Explainability Prototype	72
AI Transparency and Explainability Technical Guidance	80

Foreword



As AI makes itself even more pervasive and ambient in our lives, AI will be a necessity for companies to compete and stay relevant in today's digital economy. However, the proliferation of AI systems also introduces risks that need to be addressed. Today, consumers place responsibility for the way AI is used firmly at the door of the companies that deploy and develop it. Governments in parallel also respond, some with comprehensive guidance while others are planning for regulations on AI.

In this area, the urgency lies in the creation of tangible solutions to benefit consumers. How do we support AI system owners and AI solution providers roll out explainable, transparent, fair and human-centric AI? Singapore has been an early mover for establishing detailed voluntary guidelines to govern the development and use of AI for industry. Having introduced Asia's first Model AI Governance Framework (Model Framework) in 2019, and later a companion Implementation and Self-Assessment Guide for Organizations (ISAGO), this year, Singapore took yet another step forward by launching A.I. Verify, the world's first AI Governance Testing Framework and Toolkit to promote transparency between companies and their stakeholders.

As Singapore strives to translate high level AI ethics principles into practical implementations that benefit businesses and end users, we will need to continue to leverage industry's feedback and insights to inform, test and validate policy development. Initiatives like the Open Loop program foster closer collaboration between industry and policymakers and enable co-creation of both policies and tangible solutions.

It is in the spirit of collaborative experimentation and learning that IMDA embarked on this policy prototyping exercise on AI transparency and explainability with Meta. One of the key policy insights and recommendations of this report is important to help connect the dots - complementing policy guidance with user experience design archetypes that translate policy principles into user experiences and interactions.

AI presents complex challenges to both regulators and industry alike, which require a collective will for deeper collaboration and engagement. We are only just starting our journey. We look forward to future editions of the Open Loop program to help build and implement trustworthy AI for all.

Yeong Zee Kin

Assistant Chief Executive (Data Protection and Innovation), Infocomm Media Development Authority of Singapore and Deputy Commissioner, Personal Data Protection Commission

Foreword



Transparency and explainability are fundamental to responsibly developing AI. The OECD identifies transparency and explainability as one of its core principles, stating in its AI Recommendation that “AI actors should commit to transparency and responsible disclosure regarding AI systems... [by providing] meaningful information [that is] appropriate to the context, and consistent with the state of art.” At Meta, we agree, and these tenets are core aspects of our five pillars of Responsible AI.

And we are putting them into action by building AI transparency and explainability into our internal tools and frameworks. Our dedicated cross-disciplinary Responsible AI (RAI) team has worked closely with academia, civil society, governments, and other industry partners on several transparency and explainability projects, including AI System cards - a tool designed to provide insights regarding AI systems’ underlying architecture to better explain how it operates. Similarly, the WAIST (Why Am I Seeing This) tool offers insights into why people see certain content, including certain ads, in their News Feed and allows them to influence the selection of content provided to them, ensuring they can better understand and control the content they’re shown.

Beyond our own practices, we’ve worked to foster the responsible development of AI systems in the open source community. For example, we released Captum, an extensible library for model interpretability built on PyTorch that helps ML researchers and developers more easily implement interpretability algorithms by helping them identify which features contribute to a model’s output. And Meta’s recently published “People-Centric Approaches to AI Explainability” report features the RAI team’s draft AI Explainability Framework, which provides guidance on how to design and develop explainability experiences in AI-powered products.

External collaboration is intrinsic to our Responsible AI efforts. The Open Loop project described in this report co-developed and tested a policy prototype on AI transparency and explainability based on Singapore’s Model AI Governance Framework and its Implementation and Self-Assessment Guide for Organizations (ISAGO). This project leveraged policy prototyping and experimental governance methodologies to generate empirical and evidence-based input to equip startups in the APAC region with the know-how, tools, and techniques they need to responsibly build and implement AI explainability in practice. This would not have been possible without the tireless work and leadership of Singapore’s Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC).

Approaching policy development in an experimental and evidence-based manner enables legislators and regulators to systematically assess the impacts of their proposals on people and businesses. This allows them to better understand how those proposals resonate with the real world before becoming actual laws and regulations. Our hope is that policymakers and stakeholders around the world benefit from the learnings of Open Loop and initiate similar prototyping initiatives, embracing this innovative and collaborative way to develop laws and policies.

Erin Egan

VP & Chief Privacy Officer, Policy, Meta



xecutive summary

Open Loop is a global program, supported by Metaⁱⁱ, that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies. Open Loop builds on the collaboration and contributions of a consortium composed of regulators, governments, technology businesses, academics and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rulemaking processes in the field of technology policy.

Explainable artificial intelligence (XAI) - in its current state - is primarily used for internal audiences, rather than external ones. It tends to be better served as an internal resource for engineers and developers who leverage explainability to identify errors and debug AI and machine learning (ML) models, rather than for providing explanations to other users employing these models, or to the end users affected by these models. There is still little understanding of how AI explainability can be built for stakeholders outside of the engineering and tech development realms.

In this Open Loop program, through its experimental and participatory approach, we sought to address this gap by including a wider variety of perspectives into the research of AI transparency and explainability (T&E). We did so by exploring how to build AI explainability for a range of use cases and stakeholders in a more holistic and comprehensive way.

Methodology

We began the Open Loop program by ensuring that our participants understood and were able to operationalize AI explainability as a multidimensional concept. We did this by asking the companies to define the audience (who the explanation is aimed at), the context in which the explanation is provided, purpose (the goals that the explanation is seeking to achieve), and content (what will the explanation include) of the XAI solutions they were going to build. This enabled participants to shape and map their

This report presents our findings and recommendations to policymakers. In partnership with Singapore's Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), AI Singapore, Aicadium, TTC Labs and Craig Walker, we worked with 12 companies across the APAC region to co-develop and test a policy prototype on AI transparency & explainability based on Singapore's Model AI Governance Framework (MF) as well as its Implementation and Self-Assessment Guide for Organizations (ISAGO).ⁱⁱⁱ To do so, we designed and deployed the program to achieve the following **goals**:

- Test Singapore's AI governance framework and accompanying guide (MF and ISAGO) in the field of AI T&E, with a focus on AI explainability, for policy clarity, effectiveness and actionability.
- Make recommendations to improve specific XAI elements of Singapore's AI governance framework and accompanying guide, and contribute to their wider adoption.
- Provide clarity and guidance on how companies can develop explanations for how their specific products and services leverage AI/ML to produce decisions, recommendations or predictions (XAI solutions).
- Showcase best XAI practices.
- Offer evidence-based recommendations for AI T&E in the APAC region.

explanations to specific use cases, which we called "explainability scenarios". Through this scenario-based approach, participants personalized and tailored their explanations to their own specific AI products and services, as well as to their business models and needs.

We then observed and documented how participants built their XAI solutions according to the explainability scenarios by following a mobile ethnography approach - a qualitative

research method conducted remotely through a smartphone or desktop application. Using that approach, we engaged with the participants through what we called "missions", a series of regular touchpoints where we asked our participants to answer different types of questions, perform small tasks, upload short videos, or meet with us for live conversations or

interviews about their XAI solutions. Through this methodological approach, we captured the experience of participants in receiving, handling and following the policy prototype, and learned how they made use of it to build and deploy AI explainability solutions in practice, in the context of their everyday business lives.

AI Transparency & Explainability

A Policy Prototyping Experiment

Methods

Mobile Ethnography
& Explainability Scenarios

Participants

Bukalapak (Indonesia)	Halosis (Singapore)	Nodeflux (Indonesia)	Traveloka (Indonesia)
Deloitte (Singapore)	Meta (Global)	Trabble (Singapore)	Qiscus (Singapore)
Evercomm (Singapore)	Ngee Ann Polytechnic (Singapore)	Travelflan (Hong Kong)	QSearch (Taiwan)

Phases

1. Foundational Phase

Analytical Perspectives:

- Policy Clarity
- Policy Effectiveness
- Policy Actionability

2. Procedural Phase

Trade-offs:

- T&E vs Security
- T&E vs Effectiveness/
Accuracy
- T&E vs IP Disclosure
- T&E vs Actual
Understanding

3. Delivery Phase

Technical Considerations:

- Feasibility
- Quality
- Traceability
- Scalability

Policy Considerations:

- Range & Depth
- Human Factor

Usability Considerations:

- Visualization
- Customization
- Simplicity
- Limitations
- Flow
- User Empowerment

Recommendations

Get Practical
Get Personal
Connect the Dots
Get Creative Together
Test Experiment

During the **foundational phase**, we tested Singapore's MF and ISAGO focusing on their operational guidance regarding AI T&E. Our evaluation of the framework and accompanying guide was done by observing and documenting how the participating companies explained the inner workings of their AI systems,^{IV} and how they produced decisions, recommendations, or predictions using the

frameworks' guidance, which we summarized and converted into a policy prototype.

The evaluation and testing of the policy prototype, which guided the participants in the process of developing and deploying XAI solutions, focused on **three main analytical perspectives:**

Policy clarity

the extent to which the policy text can be meaningfully understood. A critical requirement for any law or policy is that the recipient of it can understand what is expected of them, based on the instructions a policy provides.

Policy effectiveness

the extent to which following the policy guidance enables one to meet its desired policy goals (in this case: ensuring that AI decision-making processes are explainable, transparent and fair; and that AI solutions are human-centric). A policy will be effective if it successfully enables the accomplishment of its goals.

Policy actionability

the extent to which the policy prototype equips its addressees with the means to implement its guidance. If, by reading the policy, one can readily act upon it and implement its instructions, the policy guidance is actionable.

Regarding **policy clarity**, participants found the text to be clear, accessible, and understandable. As a suggestion for further improvement, participants recommended restructuring the policy prototype in a more granular level, tailoring its guidance according to different stakeholders and specific use cases. This would help streamline the content of the policy guidance and, by customizing it to specific contexts, make it more

relevant to the specific actors designing, developing and deploying AI systems.

Concerning **policy effectiveness**, the policy prototype was deemed to be effective in two aspects. First, it raised awareness of the role and responsibility of AI developers in building trustworthy AI. Second, it provided high-level guidelines that enabled participants to iden-

tify the main risks of building and deploying AI systems, paving the path for them to design their products in a way that meets the goals of explainability, transparency and human-centricity. In order to improve the procedural and operational component of the policy guidance, participants suggested adding further details to the training of professionals on ethical AI deployment, and proposed including benchmarks and yardsticks to help them assess what is and what is not an ethical use of AI.

Regarding **policy actionability**, participants expressed doubts and anticipated implementation difficulties. They argued that it would be hard to translate the policy text into concrete outputs as the latter would require more detailed and practical instructions. To solve this issue, participants suggested mapping the policy guidance to the AI product lifecycle stages, articulating and connecting specific policy recommendations to the distinct technical steps that are involved when designing, developing and deploying an AI system, including its explainability components.

In the **procedural phase** of the Open Loop program, we observed and documented how participants built their XAI solutions at the technical (code) level. This involved asking participants to select the XAI techniques and methodologies^V that would underpin the actual building and operationalization of their explainability solutions. We also asked participants about the tensions and challenges they encountered when delving into this technical endeavor. As a result, we captured a series of situations where companies were asked to make important trade-offs, that is, reach a balance between two desirable but incompatible values and features. Companies highlighted the following **four main trade-offs**:

- **T&E vs Security** (enabling bad actors)
- **T&E vs Effectiveness/Accuracy**
- **T&E vs Disclosure of Potential IP Issues**
- **T&E vs Meaningfulness and Actual Understanding**

As an additional challenge, participants also discussed the criteria upon which an explanation should be required, and the quality that such explanation should have. This led them to reflect on the “principle of equivalence”, which suggests that the same standards of disclosure for human-driven decisions should be applied to decisions that have been made or augmented by an AI system.

Throughout this phase, we supported companies with a comprehensive technical assistance package, which included dedicated mentoring sessions, the use of a machine learning operations (MLOps) platform, and a comprehensive technical guidance toolkit that gave participants an overview of the latest AI explainability techniques, along with examples and illustrations.^{VI}

Finally, in the **delivery phase**, companies were asked to build an interface design for their AI explainability solution and present a correspondent communication strategy by defining where, when and how to communicate the explanation externally. This involved integrating the explanation in their product or service flow, and testing it with representatives of the audience category that the XAI solution was directed towards.

Beyond and through the goal of explainability, the participants reported **three additional goals** that they sought to accomplish in the process of developing their XAI solutions:

- Enhance the overall trust in AI/ML technology;
- Improve and refine the products or services to which the XAI solutions apply;
- “Get the record straight” regarding AI and its actual capabilities and limitations.

Still in this phase, participants shared a number of important technical, policy and usability considerations when tasked with building and delivering their AI explanations. In terms of usability and user experience, participants recommended AI explanations to be prominently visual; tailored to specific audiences; simple

Recommendations

but not overly simplified; cognizant and open about its own limitations; seamlessly integrated into the product or service flow; and designed

in a way that empowers and provides users with control options over the decision and recommendations produced by the AI/ML systems.

Based on the results of this Open Loop program, and the feedback received from its participating companies, we advise policymakers dealing with the question of how to regulate AI Transparency and Explainability to take the following recommendations into account:

**Get practical**

develop best practices on assessing the added value of XAI for companies and calculating its estimated implementation cost

Given the uncertainty regarding the return on investment of XAI, along with its implementation costs, there is a need for the development of best practices to assess the added value of XAI for the company and its users, along with reliable approaches to calculate the overall cost that the implementation of XAI solutions will represent to its developers. Policymakers can incentivize the development of codes of practice and technical guidance with specific examples of such value estimation practices and calculating approaches. This would then help the industry plan for and prepare their journey towards AI explainability, doing it so in a more confident and well-informed manner.

**Get personal**

make XAI policy guidance more personalized and context-relevant

Policymakers can also further tailor AI policy guidance to specific types of companies, stakeholders and areas of activity. Being more explicit and granular about whom the policy is addressed to in the first place, and drafting policy guidance in a way that relates and maps to the operational day-to-day company practices, could help ensure that the policy guidance is unpacked at the right layer in the company, while increasing its overall adoption and use.

**Connect the dots**

create new or leverage existing toolkits, certifications and educational training modules to ensure the practical implementation of XAI policy goals

Toolkits, certifications and educational training should complement existing and forthcoming policy guidance by going deeper into what it actually and practically means to design, develop, deploy, and explain AI systems. Policymakers can therefore use this opportunity to “connect the dots”, bridging the gap between normative guidance and practical implementation. When AI policy frameworks and regulatory guidance are connected to practical resources, companies will have a more concrete idea of the gaps they need to fill in terms of human and technical resources, as well as skills and competences.



Get creative together

*explore new interactive ways to co-create and disseminate policy,
and increase public-private collaboration*

Policymakers are encouraged to collaborate with private sector companies to conceptualize new ways to formulate and implement tech policies. This can include new processes, tools, and practices for policy co-design and development, like citizen participation, strategic foresight, crowdsourcing, and collaborative experimentation; and novel ways to disseminate policy findings, insights, and recommendations in more experiential formats, such as use case compilations, dashboards, webinars, and podcasts.



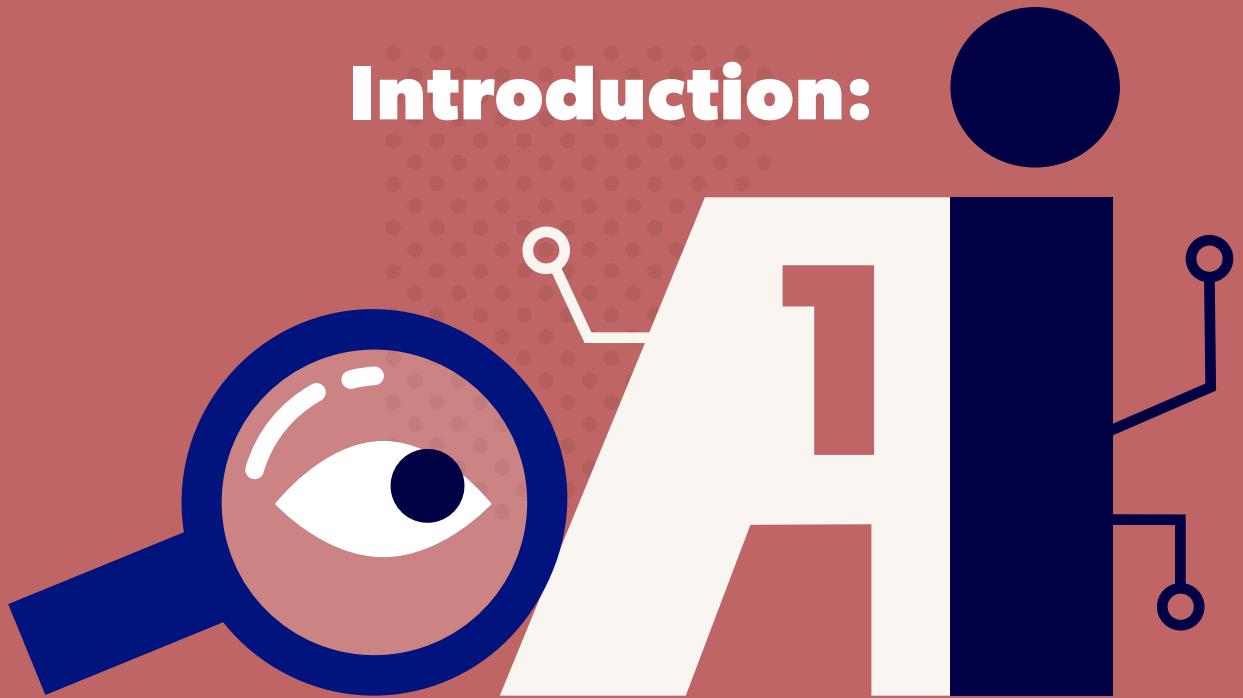
Test and experiment

demonstrate the value and realize the potential of policy experimentation

The vast majority of the participants found that testing policy ideas around AI governance is an important endeavor as it can help policymakers understand the challenges that companies may encounter when asked to follow its guidance in terms of technical feasibility and business viability.

Sandboxes and prototyping programs have the potential to shape policy and inform future laws and other governance instruments in a truly evidence-based way; but, for that to happen, policymakers need to deploy them more frequently, and assess their impact more consistently. Open Loop is a step in that direction.

Introduction:



**Transparency &
Explainability**

Understanding the concepts and their relevance

This Open Loop program focuses on the topic of **AIT&E**^{VII} which has been identified in multiple policy documents as one of the most relevant ethical principles guiding the development of AI. In fact, the OECD – which provided the first

intergovernmental standard for trustworthy AI – clustered transparency with explainability into one of its principles for the responsible stewardship of trustworthy AI.¹

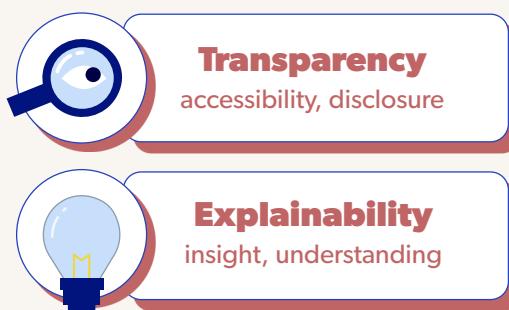
Transparency and explainability

(Principle 1.3 of the OECD, Recommendation of the Council on Artificial Intelligence)

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

- *to foster a general understanding of AI systems*
- *to make stakeholders aware of their interactions with AI systems, including in the workplace*
- *to enable those affected by an AI system to understand the outcome*
- *to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision*

In the context of this Open Loop program, we defined transparency as the ability for an outsider to access the algorithm(s), (training) data and/or learned model, while explainability is defined as an indication of the ability of an actor to understand or gain insight into how and AI system produces a given output.^{VIII} We thus argue that transparency is a prerequisite for the understanding of how AI/ML^{IX} works, while explainability entails the actual understanding of that process.



Explainability looks at the human ability to understand how and why algorithmic models (namely the ones leveraging ML) produce a given output. As Vilone and Longo put it, “[e]xplanations must make the causal relationships between the inputs and the model’s predictions explicit, especially when these relationships are not evident to endusers”.² Thus, explainability refers to enabling a particular stakeholder to understand the rationality or logic behind model outcomes.³ It can be defined as providing an overall idea of how a model functions in order to validate whether the model meets the purpose for which it was built in the first place. In this way, explainability serves specific aims or goals, including the fulfillment of particular stakeholder’s “explanatory needs” that emerge in a particular context (or use case).⁴ Ideally, explainability lets humans and ML mod-

els together arrive at better decisions than neither of them could have made in isolation.⁵

In the context of serving those explanatory needs and enabling an ideally symbiotic relationship between humans and technology, Singapore's IMDA/PDPC outlines that explainability serves to ensure that automated and algorithmic decisions and any associated data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.⁶ In this sense, explainability assumes a critical role in "[...]enabling] society as a whole to gain greater understanding about the benefits and drawbacks of AI systems."⁷

Increasing societal relevance and technical advancements

AI/ML technologies are assuming an increasingly prominent role in the development of business activities with critically important impacts on individuals and society, namely in areas such as healthcare, employment, finance, and the administration of financial justice, to name a few.⁸

This underlines the importance of empowering a greater variety of stakeholders to understand how AI/ML systems reach their decisions and recommendations,⁹ ranging from policymakers and regulators, to academics and experts and, ultimately, to the individual end users themselves.

To that effect, a number of techniques and practices aimed at making "black boxes" understandable to a wider range of stakeholders are emerging at an increasingly rapid pace

Explainability as a vehicle for Responsible AI and as a target of

Ethical AI and Responsible AI, as emerging disciplines of theory and practice, are increasing awareness about the need to design AI in a manner that operates responsibly and meets stakeholder's expectations.¹³ A wide variety of different policy stakeholders, from governments and international organizations to academics and civil society representatives, have contributed to these disciplines by publishing principles, frameworks, and recommendations, codifying a set of practices that guard against the use of biased data or algorithms, help maintain user trust and individual privacy, foster the account-

One of the main objectives of this Open Loop program is to advance new practices to explain (and meaningfully understand) how AI/ML systems reach their decisions, recommendations or predictions. For that reason, the primary focus of the program is not on transparency, but on AI explainability and the development and deployment of XAI solutions. In that context, we adopted a use-case approach to define explainability, deconstructing it in its most relevant elements (audience, context, purpose, etc) and procedures (technical, code-based, and design interface) in order to better understand and operationalize the concept.

within the engineering realm, and particularly when addressing advanced ML models.¹⁰ Important technical advancements have been made to help understand ML model behavior and automated decision-making (ADM) processes. These include: uncovering the specific weighting and importance of distinct features of a model and their correspondent effect in the overall decision produced by the ML model; generating a simpler ML model that can be more easily apprehended by humans; or providing context through counterfactual explanations - a specific class of explanation that provides a link between what could have happened had input to a model been changed in a particular way.¹¹ Such methods typically span from techniques explaining an entire model to localized techniques that explain predictions from individual instances.¹²

ability of AI actors, and ensure that automated decisions are justified and explainable.¹⁴ The voluntary adoption of Responsible AI practices, namely the ones regarding T&E, is increasingly being fostered by regulatory and operational guidance efforts across the globe, such as through Singapore's Model Framework. In effect, Singapore's IMDA/PDPC launched their Model AI Governance Framework (MF), with participation of the World Economic Forum, the European Commission's High-Level Expert Group and the OECD Expert Group on AI, as a "unique contribution to the global

⁵ Bhatt et al. 2020b (p.1-2) ⁶ IMDA and PDPC 2020 (p.64) ⁷ ICO 2018 (p.6) ⁸ Mohseni, Zarei and Zagan 2020; Wang, Xiong and Olya 2020; Ananny and Crawford 2018; ICO 2018
⁹ Bhatt et al. 2020a,b ¹⁰ Selbst and Barocas 2018 ¹¹ Verma, Dickerson and Hines 2020; Wachter, Mittelstadt and Russell 2017 ¹² Guidotti et al. 2018 ¹³ Wang, Xiong and Olya 2020
¹⁴ Accenture 2021

discourse on AI ethics by translating ethical principles into practical recommendations that organizations could readily adopt to deploy AI responsibly.”¹⁵

Moving from voluntary adoption to legal compliance, companies are adopting explainability practices in order to help ensure that their products and services are aligned with laws and regulations. The European Union’s General Data Protection Regulation (GDPR), for instance, holds multiple references to automated processing of personal data and the use of automated decision-making. It requires personal data to be processed in a transparent way, and - in certain circumstances - it sets out the right to information of the existence of ADM, including purposeful information about its rationale, meaning, and consequences (Art. 13-15). For some academics, GDPR actually enshrines a right to explanation:

“Article 22 of the General Data Protection Regulation (GDPR) sets out the rights and obligations of the use of automated decision making. Noticeably, it introduces the right of explanation by giving individuals the right to obtain an explanation of the inference/s automatically produced by a model, confront and challenge an associated recommendation, particularly when it might negatively affect an individual legally, financially, mentally or physically.”¹⁶

Such requirements relating to automated processing of personal data and the use of ADM are not unique to the European Union, and also feature in APAC privacy laws such as the Philippines’ Data Privacy Act (DPA) and its Implementing Rules and Regulations (IRRs). The DPA and its IRRs provide data subjects with a right to be informed about the existence of ADM or profiling (IRR Rule VIII, Section 34(a)(1)), and also require data controllers to notify the National Privacy Commission (NPC) of automated processing operations which are the sole basis of making decisions that would significantly affect data subjects (IRR Rule XI, Section 46(b)).

One of the reasons behind this regulatory focus, as discussed in the relevant literature,¹⁷ is the need to address the trust deficit between companies and end users. Simply put: an AI system that makes high stakes decisions that are normally made by a human needs to be able to account for how this decision came about. Singapore’s Model Framework, albeit not a legally binding regulation but a voluntary guide for companies, emphasizes that “organisations should ensure that AI decision-making processes are explainable, transparent and fair, while AI solutions should be human-centric.”¹⁸ In order to help achieve these goals, IMDA/PDPC provides a framework that helps organizations adopt accountability mechanisms in data management and protection, including clear guidance on explainability practices and procedures.



**ingapore
AI Explainability
Policy Prototyping
Program**

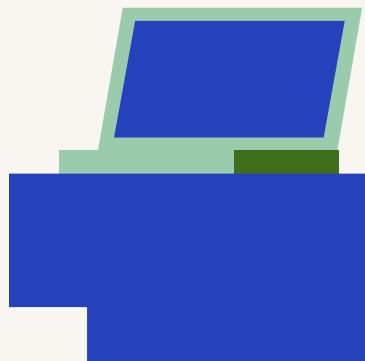
Project overview

We partnered with Singapore's IMDA and PDPC, alongside AI Singapore, Aicadum, TTC Labs and Craig Walker, to co-develop and test a policy prototype on AI T&E based on Singapore's MF and the Implementation and Self-Assessment Guide for Organizations (ISAGO). These frameworks were created to encourage organizations to take an ethical approach when deploying AI technologies. The Model Framework is sector and technology-agnostic, and translates ethical principles around explainability,

transparency, fairness, safety and human-centricity into implementable practices. The ISAGO was developed in collaboration with the World Economic Forum Centre for the Fourth Industrial Revolution (WEF C4IR), to help organizations assess how well their AI governance practices align with the Model Framework. It provides an extensive list of useful industry examples and practices to help organizations implement the Model Framework.^x

Singapore's Model AI Governance Framework: Guiding Principles

The Model Framework is based on two high-level guiding principles that promote trust in AI and understanding of the use of AI technologies.



- A** Organizations using AI in decision-making should ensure that the decision-making process is **explainable, transparent** and **fair**.

Although perfect explainability, transparency and fairness are impossible to attain, organizations should strive to ensure that their use or application of AI is undertaken in a manner that reflects the objectives of these principles as far as possible. This helps build trust and confidence in AI.

- B** AI solutions should be **human-centric**.

As AI is used to amplify human capabilities, the protection of the interests of human beings, including their **well-being** and **safety**, should be the primary considerations in the design, development and deployment of AI.

We tested and evaluated the operational guidance on AI transparency and explainability of Singapore's MF and ISAGO frameworks by observing and documenting how 12 AI companies from the APAC region developed explanations for how their specific products and services leverage AI to produce decisions, recommendations, or predictions (XAI solutions). Participants were asked to build those solutions based on the frameworks' guidance, which we summarized and converted into a policy proto-

type.^{xi} The latter was then evaluated in terms of its clarity, effectiveness and actionability.

We structured the program into three phases - foundational, procedural and delivery and asked each of the companies to personalize and tailor their explanations according to a specific scenario, defined in terms of audience, context, purpose and content of the explainability solutions they were going to build.

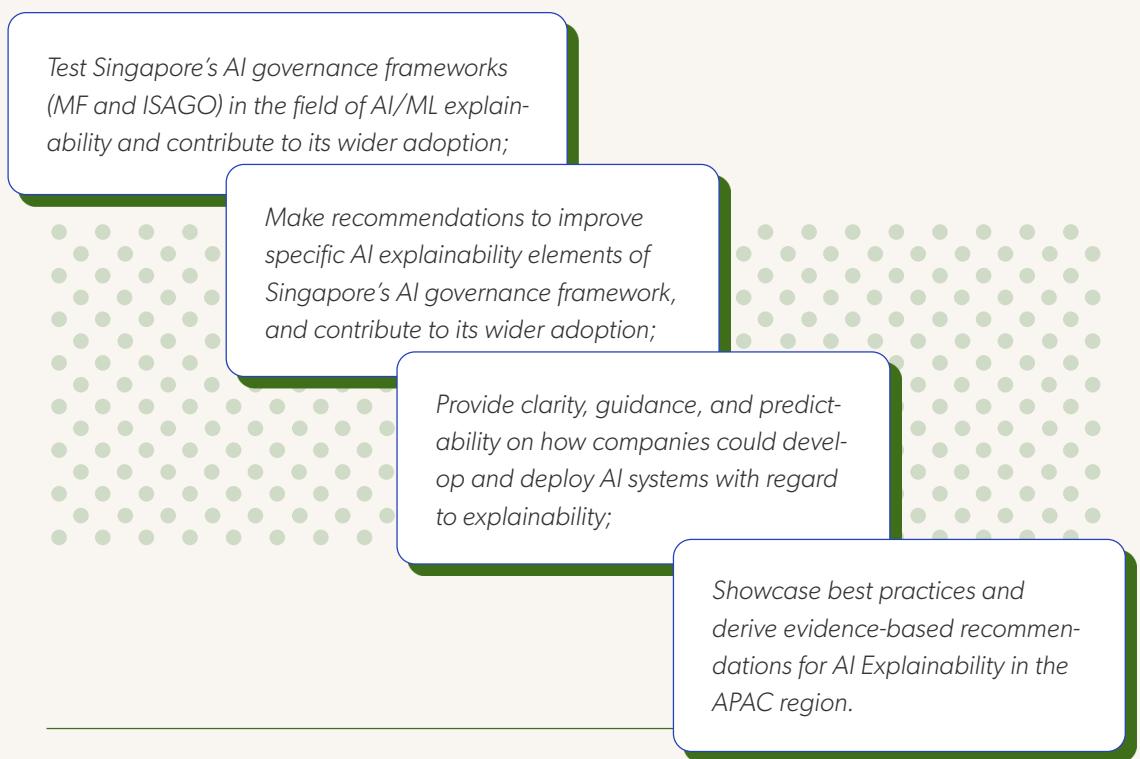
Project goals

Explainable AI – *in its current state* – is primarily used for internal audiences, rather than external ones. It tends to be better served as an internal resource for engineers and developers who leverage explainability to identify errors and debug the model themselves, rather than for providing explanations to end users affected by the models.¹⁹

There is still little understanding of how AI explainability can be built for stakeholders outside engineering and tech development realms. The

Open Loop program, through its experimental and participatory approach, aims to address this gap by including a wider variety of perspectives into the research of explainability. We did so by adopting a more holistic and comprehensive approach to explainability, exploring how to build XAI for a variety of use cases and stakeholders.^{xii}

With such a holistic perspective in mind, the Open Loop program was designed and deployed to achieve the following goals:



This was a collaborative effort, done in close partnership with Singapore's IMDA/PDPC, and alongside 12 companies across APAC which - through our methodology - tested Singapore's AI Governance Frameworks guidance on explainability. To assist the participating companies with this, we enlisted the participation of four technical partners and advisors to our program, AI Singapore, Aicaduum, TTC Labs and Craig Walker.

Partners



Singapore's InfoComm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC) develops and regulates the infocomm and media sectors. Through the PDPC, which is part of the IMDA, it also promotes and regulates data protection in Singapore.^{XIII} As an Open Loop program partner, IMDA joined our main program touchpoints, provided essential inputs in its various phases, and shared its regulatory vision and policy perspectives on the topic of AI explainability as addressed in their MF/ISAGO frameworks, which helped us evaluate and test its relevant policy provisions.



AI Singapore^{XIV} is a national AI program launched by the National Research Foundation to anchor deep national capabilities in AI thereby creating social and economic impacts, grow the local talent, build an AI ecosystem, and put Singapore on the world map.



Aicadium^{XV} is a Singapore-headquartered global technology company dedicated to creating and scaling AI solutions by leveraging deep expertise and a common machine learning platform. Aicadium partners with companies to build and operationalize impactful end-to-end AI solutions across a wide variety of industries and use cases.

Aicadium and AI Singapore provided technical assistance during the procedural phase to the program participants, helping them develop their AI explainability solutions. They hosted individualized consultations (e.g., in the form of mentoring hours) with program participants.



TTC Labs^{XVI} is a co-creation lab that advances the user experience around data. Initiated and supported by Meta, TTC Labs drives collaboration between policymakers, privacy experts and technologists through design thinking. Its vision is to create meaningful experiences between people and data that are sustainable and equitable for all.



Craig Walker^{XVII} designs and researches for the world's leading organizations, working in the technology, financial services, built environment and infrastructure sectors. Their multidisciplinary team of craft-based designers applies their expertise and experience to help clients create new opportunities and solve challenges.

TTC Labs and Craig Walker provided technical assistance to the program participants during the delivery phase, helping them present and communicate their XAI solutions. TTC Labs provided participants with a storytelling template slidedeck to help structure their presentations, along with coaching on how to deliver a pitch presentation. Craig Walker delivered insight talks to the participants on AI explainability and user design implications, further supporting the presentation and communication of their XAI solutions during the final phase of the program.

Participating companies

Key to our prototyping program was the active involvement of participating companies that not only applied and tested the policy prototype on their own selected AI products and services, but also engaged in the development of AI explainability solutions based on the guidance provided through the policy prototype. Mostly recruited through an existing accelerator partnership between Meta and IMDA, the cohort of participating companies came from a variety of different APAC countries and operated in different sectors. This geographical and sectoral diversity enabled rich qualitative insights as to how explainable AI could, and should, be achieved across different contexts, uncovering a variety of different needs and use cases. We tested the policy prototype under the lens of both startups and established companies in order to understand and improve Singapore's Model Framework for companies of

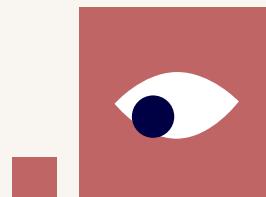
different sizes and different internal human and technical resources. All the participating companies were able to commit to and stay motivated throughout an intensive half-year journey of prototyping AI explainability around their AI/ML products and services, and in the middle of the Covid-19 pandemic.

Our participating companies cover a wide range of industries, coming from the field of Education, Travel & Tourism, Energy, E-commerce, Consulting, Information & Communication Technology, Social Media & Networking services. While this represents just a part of the industries that make use of AI in their products and services, this is a diversified and representative set of sectors that are currently building and deploying AI products and services in the APAC region.

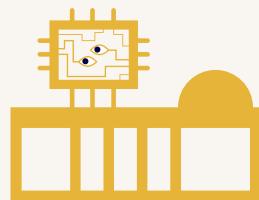
Participating companies' selection criteria



Geography



Size



Sector



Motivation for AI Explainability

The points of contact that represented the companies throughout our program also encompassed a diversified set of backgrounds: technical and engineering (as CEOs or CTOs, data analysts, data scientists, software developers, AI leads, etc), business development or prod-

uct and operations management (go-to-market leads, marketing or product marketing managers, etc), UX (user experience leads), and to a lesser extent dedicated policy and compliance or research positions (corporate affairs leads, compliance and public policy officers).

Overview of Participating companies

Total participating companies in the program: 12

Deloitte.

(Singapore)

Firm tasked with developing digital products/platforms to support Deloitte's core business in audit provision, consulting, financial advisory, risk advisory, tax and legal services.

Chosen AI application

Tool that predicts customer churn (risk sensing). Customer churn rate is the percentage of customers that stop using a given company's product or service.

Industry

Consulting/professional services

bukalapak

(Indonesia)

One of the largest e-commerce companies in Indonesia focusing on empowering local Small-Medium Enterprises (SME). Bukalapak provides AI services and solutions to domains such as shopping experience (personalization and recommendation, search), cybersecurity (scam prevention), investment (portfolio asset allocation).

Chosen AI application

Feature that assigns a scam-score to sellers in the e-marketplace used for filtering out potential scammers before incidents happen.

Industry

E-commerce

EVERCOMM

(Singapore)

An energy AI company focusing on analyzing and improving the energy consumption of clients (government and enterprises), and thus helping them transition to sustainable operations through better decision making (end-to-end energy management solutions for enterprise).

Chosen AI application

AI-enabled digital assistant combined with cost-effective IoT technology that addresses the digitization challenges faced by SMEs in the commercial and industrial industry

Industry

Energy efficiency



A real time communication (RTC) technology company/multichannel conversational platform specializing in chat and call technologies for mobile and websites, and aiming at enabling communication across any application.

Chosen AI application

Computer-vision-based intelligence videoanalytics that translates raw image and video data into insightful information for clients

Industry

Communication platform provider



Virtual social media marketer and influencer marketing platform, and a Search and Analysis Engine for Social Data. Qsearch system is used to understand social attitudes and provide analyses of them for their clients' use, both in the business and policy sector.

Chosen AI application

Feature that recommends influencers to brands through audience analysis

Industry

Social Media & Marketing



Southeast Asia's lifestyle Superapp that enables users to discover and purchase a wide range of travel, local services, and financial services products. Traveloka's comprehensive product portfolio includes transport booking services such as flight tickets, bus, trains, car rental, airport transfer, as well as access to the largest accommodation inventory in Southeast Asia, including hotels, apartments, guest houses, homestays, resorts, and villas.

Leading online travel aggregator (OTA) company across Asia Pacific providing a range of products that support travel and lifestyle activities for customers, ranging from flights and accommodations to online classes, insurance and payment products.

Chosen AI application

Tool that provides customized ranking of the hotels that fit the search criteria provided by the end user.

Industry

Travel & Tourism



Company that builds technologies that help people connect, find communities and grow businesses. Meta is moving beyond 2D screens and into immersive experiences like virtual and augmented reality, helping create the next evolution of social technology.

Chosen AI application

Tool that enables people to control and prioritize posts from the friends and Pages they care about most in Facebook's News Feed. By selecting up to 30 friends and Pages to include in Favorites, their posts will appear higher in ranked News Feed and can also be viewed as a separate filter.

Industry

Social Media

nodeflux

(Indonesia)

AI Computer Vision company engaged in various Intelligence Video Analytics (IVA) solutions for their clients. Nodeflux also develops Face-Analytics and Face-Recognition systems for various deployments.

Chosen AI application

Computer-vision-based intelligence video analytics that translates raw image and video data into insightful information for clients

Industry

Computer Vision/Information Technology

TravelFlan

(Hong-Kong)

E-commerce solutions company that supports cooperation to connect product supply chains and provide customer acquisition and retention solutions. Travelflan uses AI, Machine Learning, and Big Data technologies to provide end-to-end digital solutions. Their most recent project is a chatbot feature for concierge solutions to serve clients' customers.

Chosen AI application

Feature that provides AI-driven recommendations (e.g., activities, restaurants, tours, delivery items) on Travelflan's marketplace

Industry

E-commerce

NGEE ANN POLYTECHNIC

(Singapore)

Institute of Higher Learning in Singapore, offering full-time and part-time courses for teenage students and adult learners.

Chosen AI application

AI Chatbot, called EVA, that automates and facilitates students enrollment.

Industry

Education

Halosis

(Indonesia)

E-commerce solution provider that facilitates a safe purchasing process for consumers of products sold online, and improves operational efficiency through the use of automation and virtual assistants for social media sellers.

Chosen AI application

Virtual assistant and automation for social media sellers

Industry

E-commerce

TRABBLE

(Singapore)

AI-driven company that provides a guest engagement platform as a solution for the travel and hospitality industry. It enables businesses to support travelers continuously throughout their journey (pre-trip inquiries, reservations, in-trip checking, check-out, concierge services) with chat based AI automated solutions.

Chosen AI application

Tool that leverages Natural Language Processing (NLP) to help Trabble's customers serve their guests better.

Industry

Travel & Tourism



In Conversation with AI Singapore

Open Loop interviewed our partner AI Singapore on the state of XAI in APAC and their contribution to the program. AI Singapore (AISG) is a national AI program launched by the National Research Foundation (NRF) to anchor deep national capabilities in AI thereby creating social and economic impacts, grow the local talent, build an AI ecosystem, and put Singapore on the world map.

Open Loop: What was your role in Open Loop? And what is your take on Open Loop in the context of AI and its Governance in the Asia-Pacific region (APAC)?

Since AI Singapore started in June 2017, we have executed nearly 60+ projects and deployed more than 20 AI applications into production. Based on our experience, we have created an internal governance checklist based on IMDA's MF to suit our requirements, needs, and AI project execution pace.

The Open Loop initiative complements our mission as it helps companies execute AI projects responsibly by supporting them with operational and technical guidance.

We are one of the tech mentors in Open Loop to enable the participating companies to develop their XAI solutions. We wanted to share our experience in AI Governance, especially on explainability, which is vital in building trust in AI solutions.

Open Loop: How do you perceive the state of AI Explainability in Singapore today, or across APAC?

There is still an awareness gap in AI Explainability. We encounter many project requests where the only desired outcome is an increase in productivity or revenue, with no regard to explainability. We educate companies by explaining the importance of AI Explainability from a business risk perspective.

You mentioned the AI Singapore team started to more intensively focus on the broader scope of Trustworthy AI, going beyond the questions of AI Explainability and contextualizing it. Why is that?

Businesses need to get their stakeholders to trust their AI solutions. Otherwise, stakeholders might question their AI solutions' results or refuse to use their AI solutions altogether. Explainability could be a good starting point to build trust.

Explainability helps stakeholders to "see" how the AI solutions work. It is also essential to know whether AI solutions are working in the "right" way, i.e. whether the model is free of unwanted bias and is robust against attack. This will be easier to evaluate if the model is explainable.

Can you describe those endeavors a bit more? What are your plans and activities you foresee in that area?

AI developers realize the importance of fair and Responsible AI solutions, but they lack a holistic framework that combines both qualitative and quantitative fairness assessment with software tools to help them conduct such evaluations.

For instance, it is insufficient and vague to instruct AI developers that their AI models must be fair. AI developers would need to understand the appropriate metrics to use to measure fairness, under what context some metrics would be irrelevant, and the appropriate software tools they could use to generate that assessment. The same goes for other important considerations, such as robustness. The AI developers need not just a framework, but also the accompanying software tools to assess each component in the framework to make sense of it all.

Therefore, we started developing an AI audit framework that will help developers and organizations quantitatively and qualitatively understand, evaluate, and communicate the degree to which their AI solution will deliver its intended value and its potential effects on the business that adopts the AI solution.

The framework will provide software tools, standardized tests, evaluation methods and scoring metrics to establish good industry practices. We plan to test the AI audit framework with our industry projects under our 100 Experiments program before a general public release some time in 2022.

And if you had one solution to propose, to bridge the gap between technology and regulatory innovation, what would you do?

There needs to be more communication, cooperation, and collaboration between regulators and developers. These will help regulators understand the implications of the latest technical developments to regulate such technologies appropriately and help developers understand regulators' concerns to develop the technologies in a more acceptable direction for the regulator. If these are done well, regulators could implement effective policies that will not stifle innovation, and developers could put in place AI solutions that regulators would readily accept.

Methodology

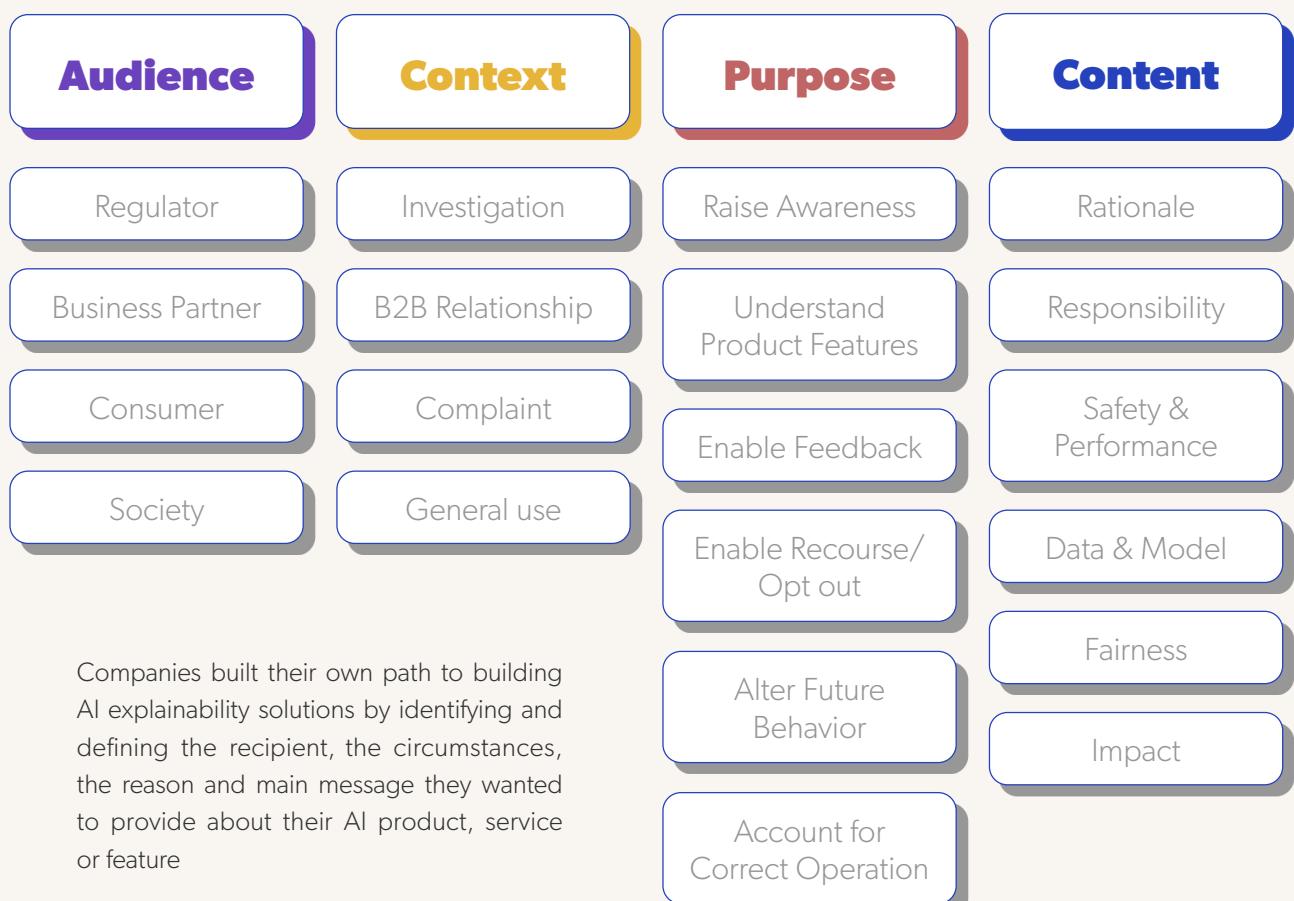
Explainability scenarios

We began the Open Loop program by ensuring that our participants understood and were able to operationalize AI explainability as a multidimensional concept. We did this by breaking down the concept of explainability into four fundamental elements: audience (whom to provide the explanation to?), context (in what context is the explanation provided?), purpose (what are the goals that the explanation is seeking to achieve?), and content (what content will the explanation include?). These four elements, each of which were in turn broken down into different categories, enabled participants to shape and map their explanations to specific use cases, which we called "explainability scenarios" (see illustration of the scenario chart below).^{xviii} These scenarios served as points of departure to help the participants build ex-

plainability solutions, forming a set of personalized pathways that companies would follow when building their explainability features. Asking companies to define the specific audience, context, purpose, and content of their explanations enabled them to tailor their XAI solutions to their own concrete AI products and services, as well as to their business models and needs.

Each of the companies were asked to build and follow two XAI scenarios. This was done to evaluate and test the XAI elements of the Singapore governance frameworks in the most granular and comprehensive way possible, looking at all the different possible contexts in which its provisions on stakeholder engagement and communication could be applied to.^{xix}

Explainability scenarios chart



Audience

refers to the recipients of the explanation

When defining and characterizing the target audiences, it is important to ask:

- Who is the target audience?
- What's their level of knowledge or expertise?
- Does the person receiving the explanation have expertise in the domain the decision is made?
- Or do they have no specialist knowledge?
- What's the understanding about their needs and expectations?
- What are their different interests in the explanation?
- What's their own interest in improving their understanding?

A balance must be struck between completeness of the explanation and the interpretability for the subject/users, i.e., information for subjects should neither be overloaded nor oversimplified.

Four main target audiences were identified:

Regulator/auditor (ext)

Includes external auditor

Business partner

Another company, client, vendor, etc.

Consumer

The user of your AI product, service or feature

Society

The public in general

Context

refers to the circumstances under which the explanation is being provided. Depending on context, the what, when, how, why and who of explanations can change dramatically.

There is no one-size-fits-all approach for explanations of AI decisions, the context in which an AI decision is made affects the importance of receiving an explanation.

The content and delivery of explanations should be tailored to their audience based on a consideration of the relevant contextual factors

The importance of an explanation of an AI decision is likely to vary depending on the person it is given to. For example, in a healthcare setting, it may be more important for a healthcare professional to receive an explanation of a decision, than for the patient, given their expertise and authority in this context.

Four different types of context were identified:

Investigation

e.g., initiated by a regulatory entity, auditor, etc.

B2B Relationship

e.g., collaboration, procurement

Complaint

e.g., from consumer

General use

explanation as part of offering your services, products, features

Purpose

refers to what the explanation is trying to achieve, that is, the objective and/or main motivation for why an AI explanation is provided.

The purpose of the explanation is different from the overall purpose of the AI system. The purpose of the explanation should be linked to the audience and context of that explanation.

Seven different types of purpose were identified:

Raise awareness

to the fact that the user is interacting with an AI system

Understand product features

e.g., collaboration, procurement

Enable feedback

from consumer, business partner, regulator

Involve Users

in improving the AI system

Enable Recourse/Opt Out

of the person or entity affected by the AI system's output

Alter Future Behavior

of the person or entity affected by the AI system's output

Account for Correct Operation

of the AI system

Content

refers to the information given to the recipients of the explanation, that is, what to focus on as information to be provided.

Based on Information Commissioner's Office (ICO) and Alan Turing Institute Explainability project.²⁰

Six main types of explanation, content-wise, were identified:

Rationale

information describing how the AI system prediction was made

Responsibility

information about who is involved in the development, management and implementation of an AI system

Safety & Performance

information about the accuracy, reliability, security and robustness of predictions and behavior of the AI system

Data & Model

information about the data training sets, algorithmic models used, etc

Fairness

information ensuring that the AI system is not unfairly biased

Impact

information about the impact that the use of an AI system and its predictions may have on individuals

Based on this set of personalized explainability scenarios, the participants completed the program by going through 3 phases: foundational, procedural, and delivery. Along this journey, and through a mobile ethnographic approach, companies provided detailed insights about their experience building and

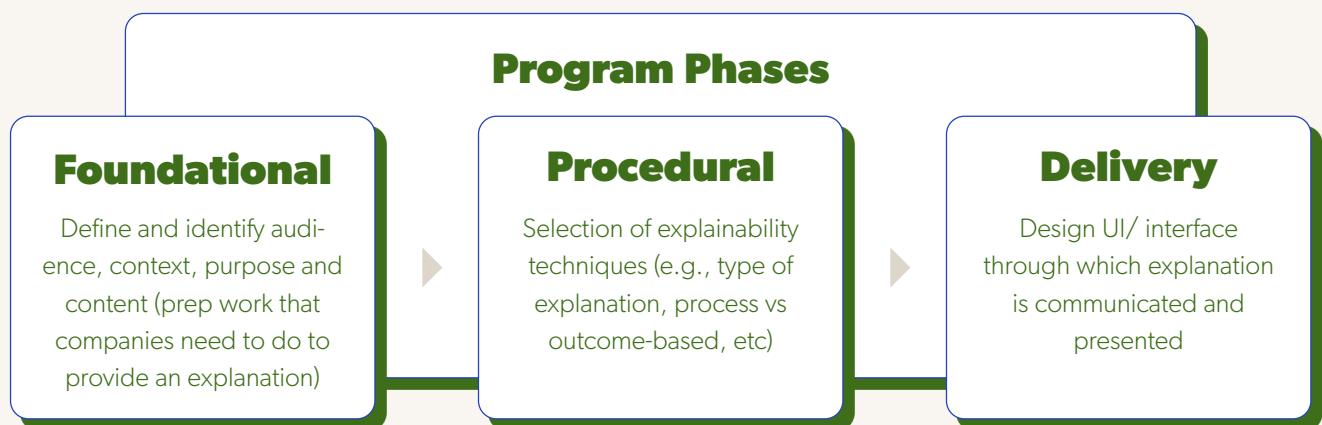
delivering the explainability solution under the guidance of the policy prototype.

Let's take a closer look at the main phases of the program and components of mobile ethnography.

Program Phases

The overall program was structured into three chronological phases: foundational, procedural, and delivery. The working assumption underlying this structure and methodology is that there is no one-size-fits-all approach for explanations of AI outputs. This is a deeply contextualized effort. The development, design, and delivery of AI explanations needs to: take

into account a number of contextual factors (amongst which are audience and purpose); select the specific XAI techniques appropriate for its use case; reflect and make trade-off decisions amongst competing or conflicting values; and choose amongst different visualization and presentation interfaces in order to be clear and meaningful.



During the foundational phase, we collected the participating companies' feedback on the specific AI transparency and explainability elements of Singapore's MF and ISAGO frameworks. We did this by asking the participants to build their XAI solutions based on the frameworks' provisions that were included in the policy prototype.

The testing and evaluation of the policy prototype, which guided the participants in the process of developing and deploying XAI solutions, focussed on three main analytical perspectives: policy clarity, effectiveness and actionability.

In the procedural phase, participants were asked to select the XAI techniques and methodologies that would underpin the actual building and operationalization of their explainability solutions. This phase focused on implementing

explainability into an algorithmic model by suggesting concrete ways of how to build explainable AI from a technological standpoint. To support companies in this technical endeavor, we provided a comprehensive technical guidance toolkit that gave participants an overview of the latest AI explainability techniques, along with examples and illustrations.^{xx} We also offered expert talks and mentoring sessions with our program's technical partners: AI Singapore, Aicadum, TTC Labs and Craig Walker.

In this phase, and to better guide participants in navigating the nuances involved in building XAI features, we also asked a number of questions regarding the value-based trade-offs and challenges they would need to address when technically constructing their XAI solutions. This enabled companies to make balanced and appropriate decisions throughout this technical endeavor.

In the delivery phase, the final one, companies were asked to build an interface design for their AI explainability solution and present a correspondent communication strategy. This task involved choosing the where, when and

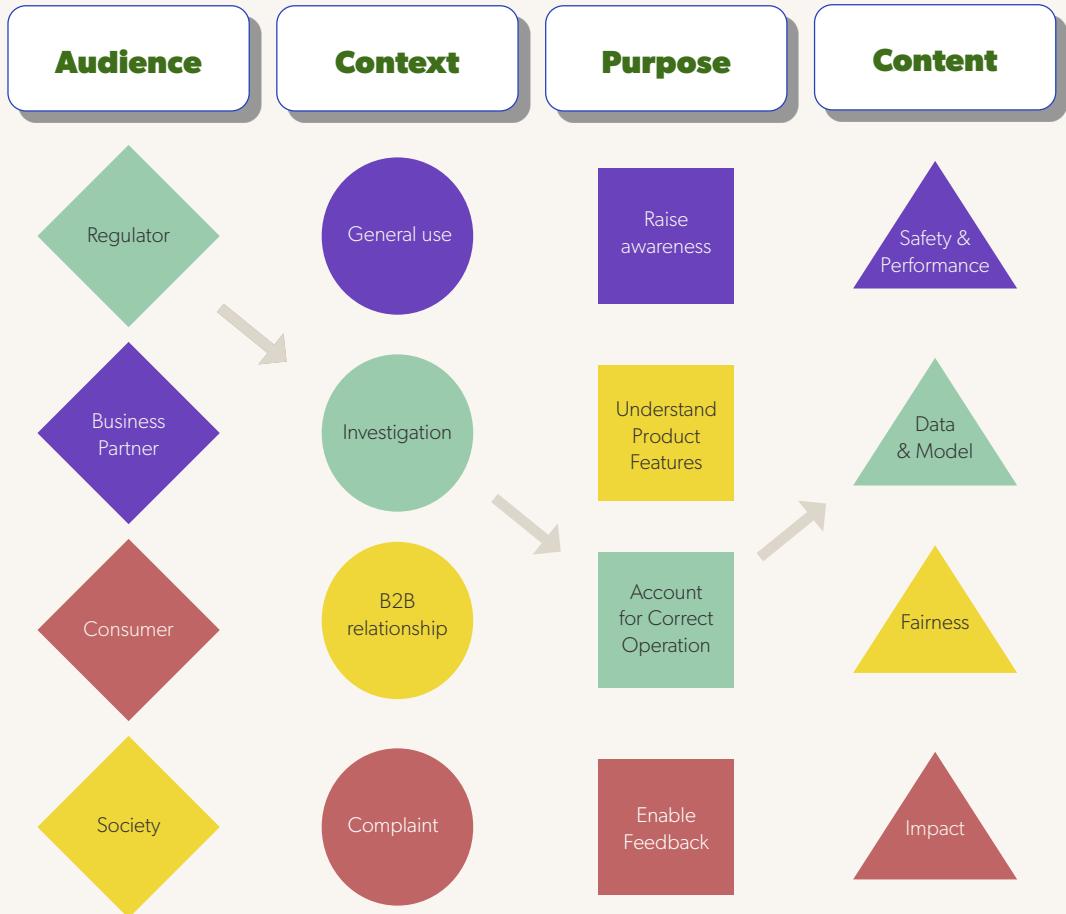
how to communicate the explanation in their product or service flow, and - ideally - to test the delivery of such explanations with representatives from the audience category that the XAI solution was directed towards.

Program Phases & Dynamic/ Personalized Scenarios

Structured into
3 chronological phases



Implemented through
a series of dynamic
scenarios built and
personalized to par-
ticipating companies
business model
and needs



**Mobile
Ethnography**
**A tool to capture
insights from
policy experience**

We observed and documented how participants built their XAI solutions according to their explainability scenarios by following a mobile ethnography approach, which is a qualitative research method that is conducted remotely through a smartphone or desktop application. It enables in-context, participant-based research that is prompted and moderated through specific "missions". These "missions" are composed of different types of

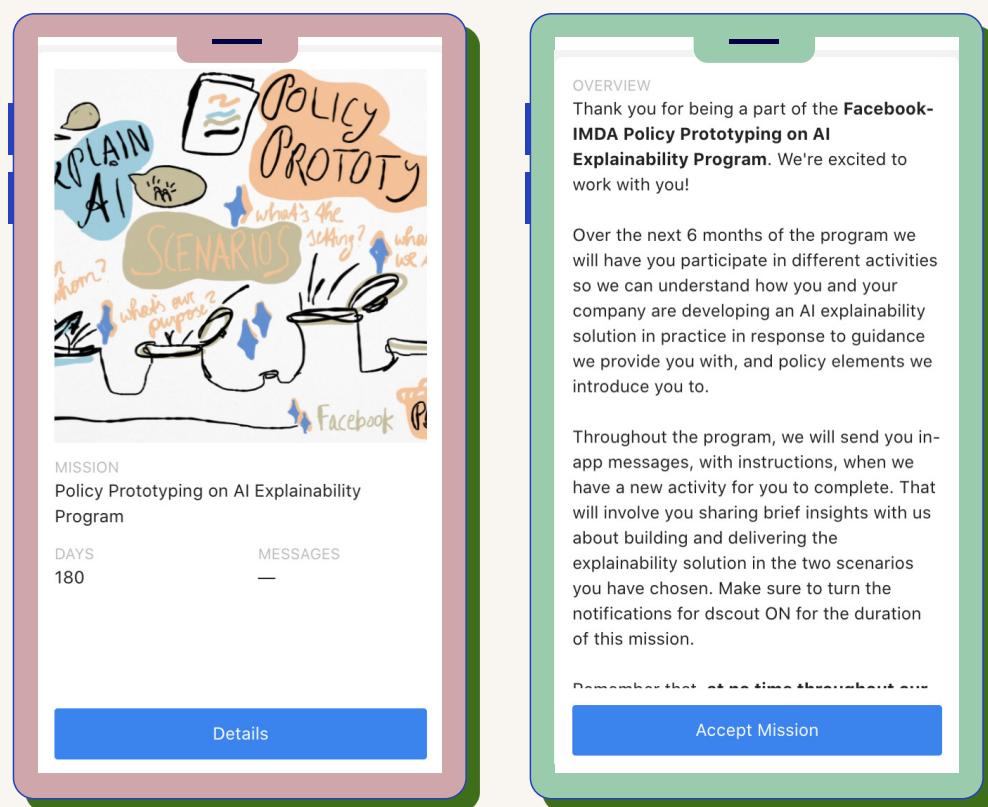
questions, small tasks, short videos, or asks for live conversations or interviews, and are used to explore a given research topic. "Missions" are also used as touchpoints to moderate the interaction between the researcher or program manager and program participants. There are a number of platform providers for mobile ethnography today,^{xxi} providing an easy-to-use and user-friendly interface for both researchers and research participants.

Mobile ethnography has been recurrently used in market research to examine in-context user experience, namely for particular products or services. We creatively applied and repurposed this approach to our policy prototyping program, leveraging it to capture the experience of participants in receiving, handling, and following a given policy framework, while learning how they made use of it in practice.

With our policy prototyping approach we wanted to be able to capture insights that came as close as possible to the actual experience of having tech companies develop their own explainability solutions while following normative provisions, and applying operational policy guidance to their product development and business model processes. Through this lightweight and user-friendly approach, we managed to understand and document how

companies received and applied the policy prototype in the context of their everyday business lives, and not isolated or detached from them. This approach highlighted the procedural and contextual nature of the explainability solutions that companies were asked to develop in accordance with the guidance from the policy prototype.

The possibility of collaborating and working remotely through mobile ethnography was particularly important given the Covid-19 pandemic and its many (physical) restrictions. Converting limitations into opportunities, this approach enabled us to improve the research design of the program, opening up new interaction and collaboration opportunities and formats via regular activities, prompts, quizzes and survey-like questions.





oundational Phase

Evaluation and Testing
of the Policy Prototype

Once we had all the Open Loop partners and participants onboarded, and the topic, goals and methodology of the program defined, we were ready to start the evaluation and testing of the policy prototype. As mentioned above, we focused on the T&E components of Singapore's MF and ISAGO frameworks. We then

broke down the policy prototype in multiple parts, evaluating and testing its content in accordance with the program phases outlined above and based on the scenarios chosen by participating companies. The evaluation and testing of the policy prototype focused on three main analytical perspectives:

Policy clarity

the extent to which the policy text can be meaningfully understood. A critical requirement for any law or policy is that the recipient of it can understand what is expected of them, based on the instructions a policy provides.

Policy effectiveness

the extent to which following the policy guidance enables one to meet its desired policy goals (in this case: ensuring that AI decision-making processes are explainable, transparent and fair; and that AI solutions are human-centric). A policy will be effective if it successfully enables the accomplishment of its goals.

Policy actionability

the extent to which the policy prototype equips its addressees with the means to implement its guidance. If, by reading the policy, one can readily act upon it and implement its instructions, the policy guidance is actionable.

Policy clarity

Policy clarity refers to the extent to which the participants understood the meaning of the text and its key concepts. There are three important contributing factors that determine the clarity and enhance the understanding of a policy prototype:

Content, Style, and Structure

Content is about what information is provided and made available. Despite specific comments on the lack of clarity in certain specific sections of the policy;^{XXII} overall, participants found the text to be generally clear, accessible, and understandable. The participants also found the language to be drafted in an informal and user-friendly manner.

"The sentences are written in a very conversational way, so it's not hard to understand what each sentence means"^{xxviii}



The companies also appreciated the fact that the document did not present information in a legalistic-style format:

"At first I thought this was going to be a standard legal [text]...but I was surprised how easy it was to digest this document."



Structure is about how and where that information is presented in the policy prototype. The way the information was structured resonated well with the participants.

"I think this document is very clear, because it first describes the goal in one short sentence and then provides further details through explanations accompanied by examples."



Just as important as what information is communicated, is what information is not communicated. The participants flagged that there were parts and concepts within the policy prototype that were not applicable to their individual cases.

Policy effectiveness

The effectiveness of a policy prototype refers to the extent that its guidance contributes to reaching the desired policy outcome. In the case of Singapore's MF and ISAGO frameworks, the de-

dual case. Guidance regarding interactions with consumers and end users, for instance, would not be applicable to B2B companies. And generic requirements for explainability disclosures would need to take into account the multiplicity of actors involved in building, developing, deploying, and monitoring AI systems. From this specific piece of feedback given by our participants, one could see the value of restructuring the policy prototype in a more granular level, tailoring its guidance according to different stakeholders and their specific use cases, and knowing what, and what not, to include.

"It is important to know who we are communicating with because that entails different ways of communicating and different levels of detail to be conveyed."



This would help streamline the content of the policy guidance and, by tailoring it to specific contexts, making it more relevant to its particular addressees. Nonetheless, and as a generic document of orientation applicable across different groups of stakeholders, there is obviously a delicate and fine balance to be struck between personalizing the policy framework to its individual addressees and keeping it broad and high level. A policy prototype cannot (and should not) cover every possible use case and correspondent audience segment. It needs to remain flexible and adaptable to an evolving variety of use cases and actors, while still relevant and useful for individual instances.

sired policy outcome coincides with its guiding principles: ensure that the AI decision-making process is explainable, transparent, and fair; and the AI solutions are human-centric.

Overall, the policy prototype was deemed to be effective in two important aspects:

1

raising awareness of the role and responsibility of AI developers in building ethical AI; and

2

providing high level guidelines that enable them to identify the main risks involved in building and deploying AI systems.

In testing the policy prototype with language taken from MF's foreword, objectives, and guiding principles, our participants reported that the document increased their awareness of AI Ethics as a set of principle-based practices, and that it drew their attention to a number of issues they hadn't thought of or encountered before.

"I think the language makes sense and incorporates a humanistic view on how to think about AI for the public good."



"Although perfect explainability, transparency and fairness are impossible to attain, organisations should strive to ensure that their use or application of AI is undertaken in a manner that reflects the objectives of these principles as far as possible. This helps build trust and confidence in AI."



"The adoption of AI should be ethical, human-centred and pragmatic."



"I am now more aware of the very real impact that AI has on both users and AI developers"



"Before this program, we never really considered AI ethics"



"[The policy prototype] raises many questions which previously I have not thought about when developing AI services."



The participants also stated that the policy prototype gave them the pointers that they needed to flag the main risks posed by the development and deployment of AI systems.

"The impact on society/ethical AI deployment depends on the groundwork/how the company addressed these topics from the start."



The participants also made a number of suggestions to further improve the effectiveness of the text of the policy prototype in achieving its desired policy goal (AI that is explainable, transparent, fair, and human-centric). One of those suggestions revolved around the need for more procedural and operational guidance, as well as an increased layer of granularity and specificity in the way it conveys its guidelines. In certain sections of the policy prototype, some participants argued that there were not enough details on "how" they could implement and operationalize ethical and trustworthy AI in practice. For that reason, participating companies recommended complementing the current references to training sessions in the prototype

with further details regarding their possible scope, objectives and curriculum. This would help companies understand the type of skills and competences that their staff would need to be trained on. Ethical training was thus seen by the participants as a foundational instrument to operationalize ethical AI.

Related to the need for a text that would be more operational in nature, the participants also recommended the inclusion of benchmarks and yardsticks through which to measure what is and what is not an ethical use of AI. The participating companies encouraged the policy guidance to go beyond the reference to principle-based goals of explainable AI, and equip companies with the resources to understand where they are in that journey, that is, to measure their progress towards achieving that goal.

"We also felt that there was no yardstick for companies to use in measuring what qualifies as an ethical use of AI."



"Are there any acceptable quantitative indicators to measure explainability, transparency, fairness?"



A number of participating companies also recommended incorporating examples of what it means to achieve and what it means to not achieve the goals of the policy prototype. Such examples would help companies gain a better understanding of the impact and consequences that would result from accomplishing the goals of the policy guidance, and the impact and consequences that the policy is trying to avoid when its goals are not met.

"It may be useful to provide some examples of AI in decision-making that was non-explainable, non-transparent and unfair, and to elaborate on the severity of such an issue/consequences of it. [...] it will certainly be useful to understand the severe consequences that, as a company, we should avoid altogether."



Examples were also seen as a great way to make "generic and broad" concepts more tangible for the participating companies. These examples could also describe (and anticipate) the "pain points" that companies should expect when engaging in the process of building and implementing AI explainability.

"Maybe can include points like: what are some of the difficulties in implementing this specific suggestion and how would you go about doing it."



A policy prototype that would include these additional elements (benchmarks, examples of both positive and negative outcomes, along with an illustration of its consequences), would provide AI developers with a more realistic and accurate sense of the costs and benefits entailed in adhering and following the policy prototype's guidance. It would also enable these companies to anticipate the repercussions on their business model from following the policy guidance.

Lastly, participants also stressed the importance of strong internal buy-in of the policy guidance from leadership teams.

"This type of document needs to be endorsed and adopted by high-position management, otherwise it is unlikely to gain traction."



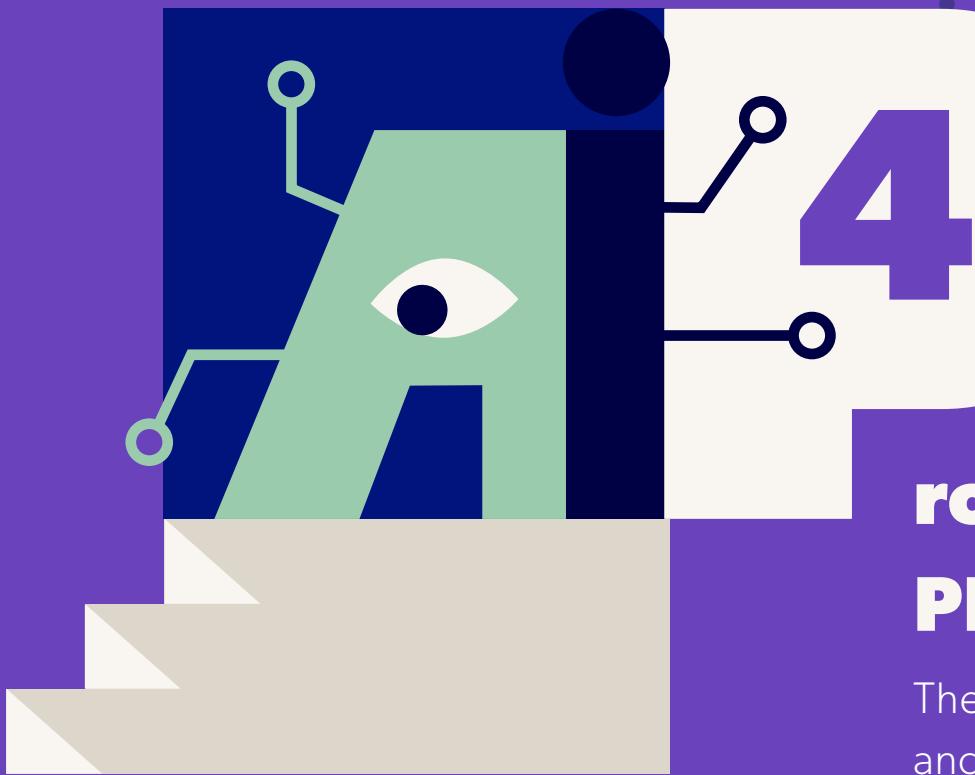
Policy actionability

The actionability of a policy prototype refers to the extent to which the policy equips its addressees with the means to implement its guidance, that is, with the resources needed to convert its guidance into actual practices. Participants expressed doubts on the actionability of the policy, and anticipated difficulties in implementing it. They argued that it would be hard to translate the policy text into concrete outputs as the latter would require more detailed and practical instructions.

"The document lacks the details to fully put its guidance into practice"



To get closer to operational mode, especially in the sections of the policy prototype that may be less precise, the participants surfaced the need for more detailed information in order to be able to act upon it, particularly if the policy should speak to non-legal functions, e.g. engineers, product designers, or AI researchers. They suggested mapping the policy guidance on AI to the AI product lifecycle stages, articulating and tailoring specific policy recommendations to the distinct technical steps that are involved when developing and deploying an AI system (training, testing, validating, etc).



procedural Phase

The technicalities
and the trade-offs
of building
AI explainability
solutions

In the procedural phase of the program, participating companies shared with us the rationale and choices behind the technical elements underpinning XAI solutions that they foresaw to build. This phase looked at the algorithms and models that were being used to build XAI, delving into the particular types of ML algorithms, datasets, and explainability techniques leveraged by our participants.^{xxiv}

In this phase we also asked participants about some of the challenges they faced when implementing T&E at the technical level, namely regarding the value-based trade-offs they encountered and had to decide on.

Technical assistance provided to the participants

Given the technically complex task at hand, we supported companies throughout this phase with a robust and comprehensive technical assistance package, which entailed:

1

Dedicated mentoring sessions and expert talks with our technical program partners Aicadium and AI Singapore.

2

Opportunity to use an end-to-end ML platform that helps companies build explainability features into their models and AI systems.^{xxv}

3

A comprehensive AI Transparency and Explainability Technical Guidance,^{xxvi} which provides an overview of AI explainability from a technical perspective, and strategies to render algorithms explainable, including supplementary techniques, and background information (e.g. tutorials, R packages, Python libraries).^{xxvii}

The vast majority of the participating companies engaged in one or various modalities of technical assistance offered to them. They qualified them as a “learning opportunity” that allowed them to “stay ahead of the game”, while helping them reflect on the progress made to date

in their XAI journey. Bedrock, the MLOps platform supporting the deployment of Responsible AI, was described as “clear” and “accessible”, helping the participants acquire a “better understanding of the process of model deployment” (Deloitte).



In conversation with Alcadium

Open Loop interviewed our partner Alcadium on their role, contributions and lessons learned from the program

What is Alcadium and what do you do?

Alcadium is a Singapore-based AI software and global technology company dedicated to creating and scaling AI solutions by leveraging deep expertise and a common machine learning platform. Through meaningful engagement and collaboration, we partner with companies to build and operationalize impactful end-to-end AI solutions across a wide variety of industries and use cases.

When and why did you decide to join the Open Loop program?

In July 2020, we joined the Open Loop APAC program as its private-sector technical assistance partner. Given our expertise and commitment to building well-governed, trustworthy AI systems, we saw this partnership with Meta and IMDA under Open Loop as a natural extension of our efforts to contribute to research that will bridge the policy and innovation gap in Responsible AI, enabling wider adoption of AI/ML explainability.

How did you go about building your MLOps platform, Bedrock?

We have taken the principles from the Singapore Model AI Framework and identified different sections to build into our machine learning platform, Bedrock. For example, we have built fairness and compliance directly into Bedrock, giving equal prominence to an algorithm's performance and its fairness and compliance score.

What was your role in the Open Loop program?

We guided enterprises in developing their XAI solutions based on a series of dynamic scenarios built and personalized to each participating company. We also mentored and guided participants in using Bedrock to explore technical options and grasping trade-offs when choosing their XAI algorithms. This was the case of Deloitte Singapore, who we helped use Bedrock to build explainability and fairness features into its machine learning models for decision making systems.

What did you learn from mentoring Open Loop participants?

Through the mentoring sessions, it was evident that the journey to Responsible AI does not end with AI ethics. For an enterprise, Responsible AI is multidimensional. There is a philosophy of how you want to be responsible in the use of AI that is expressed through your governance structures, processes and risk management policies. There is also the dimension of managing stakeholder expectations on how AI is used and its impact on users. On a day-to-day basis, there are also multiple decisions to make about how to mitigate bias, what is an

appropriate threshold for fairness and when to retrain a model. Responsible AI is about finding an integrated way of managing the AI development across the life cycle in a way that is aligned with the expressed values and ethical standards of the enterprise.

As a company sitting in the intersection of governance and innovation, we had to work with companies to grapple with the real world implications of adopting Responsible AI. For example, insisting on full transparency may expose the enterprise to greater business risks. Different users will also have varying levels of understanding and expectations about the degree of explainability they would like. One practical insight gleaned from the Open Loop participants, in the spirit of prototyping, was to explore providing explanations through repeated interactions with users, rather than a one shot approach.

Did any specific collaboration with the Open Loop participants stand out and continue beyond the program?

We are delighted to be able to continue the collaboration with one of the Open Loop participants, Nodeflux, a leading computer vision start-up in South-East Asia, to develop explainability for facial recognition and object detection. In the words of Liu Feng Yuan, Vice President of Business Development at Alcadium and former chief data scientist of GovTech Singapore, "this is pioneering work that will contribute to enhancing public safety in the use of computer vision."

Adhiguna Mahendra, PhD - Nodeflux Chief of AI Research and Product Innovation commented that "our computer vision systems and products have been implemented in a wide range of sectors such as smart cities, defence and security, banking, retail and wholesale store analysis. We realize that AI explainability will be a very important component of AI implementation in the future and we gathered valuable insights about XAI under collaboration with Alcadium during the Open Loop program. Therefore we are excited to continue this partnership to use Bedrock platform and Alcadium deep XAI know-how to strengthen the explainability of our AI solutions."

If you had one solution to propose, to bridge the gap between technology and regulatory innovation, what would that be?

Technology and regulatory innovation need not be at odds with each other. The next step is about translating Responsible AI principles into metrics that can be practically applied by data scientists, business owners and those in risk management to verify the trustworthiness of their AI solutions.

Trade-offs involved in building Transparency & Explainability

During the procedural phase we also asked our participating companies to identify and describe the tensions and challenges that they encountered when building their XAI solutions at the technical level. As a result, we captured

T&E vs Security

(enabling bad actors)

One of the trade-offs identified by our participants consisted in how T&E, in certain situations, may empower and enable bad actors to act more effectively, gaming the system and manipulating algorithms for their own purposes. To mitigate this risk, some of the companies argued that it may be necessary to keep a minimum level of opaqueness about how their algorithms operate, undermining explainability

T&E vs Effectiveness / Accuracy

Another trade-off identified by the participating companies argued that certain levels of transparency on AI models may come at the expense of their effectiveness and accuracy. Given that for modern AI methods, especially deep learning, there is often a correlation between the model's effectiveness and the difficulty in understanding it, a series of intriguing questions emerge: when does it make sense to simplify these models for human understanding at the expense of their effectiveness? In other words, when does it make sense to low-

T&E vs Disclosure of Potential IP Issues

As noted by some of our participants, full transparency of algorithms, namely disclosure of source code, raises important legal problems from intellectual property and trade secrecy perspectives, just like the disclosure of other types of proprietary information (e.g. software, patents). In effect, overly strict transparency requirements re-shift incentives to innovation that may have negative unintended consequences. Companies equated this risk with the one of security that was previously referred to. Finally, several of our companies expressed the opinion that the value of explainability should be prioritized and models be made interpretable,

a series of situations where companies were asked to make important trade-offs, that is, reach a balance between two desirable but incompatible values and features. Companies highlighted the following four main trade-offs:

in order to ensure the security of the AI system. Other participants were not so sure about the need for this compromise, and mentioned that the kind of information and level of detail required to explain how an AI model works to the end user will likely not correspond to the level of information and detail a malicious intent user would need to game or manipulate the model to its own advantage.

er accuracy for a gain in transparency? When would one prioritize receiving an explanation to the detriment of accuracy and vice-versa?

"One trade-off is the model accuracy. We know that certain models such as decision trees are able to explain the process, but the performance is not as good as those black box algorithms."^{XXXVIII}

Deloitte.

while IP rights may be minimized to the extent possible, in cases where AI systems can lead to significant impact on human safety or if they are expected to result in significant impact on human lives through their decisions, predictions or recommendations (Nodeflux, Bukalapak, Deloitte).

Achieving "[...] full transparency is not something that is easy to be done. That involves not only IP issues, but also security issues."

 **bukalapak**

T&E vs Meaningfulness and Actual Understanding

Our participating companies also argued that overly detailed T&E may not be meaningful to users and may not advance the understanding of how their data is being handled, and how decisions, recommendations, and predictions are made. As noted by our participants, extensive description of the inner logic of algorithms, which may only be understandable by experts, may not contribute to explaining to users how automated decision processes are attained. On the contrary, it may do the opposite, overwhelming and confusing them even more. Giving individuals too much information about AI systems and their outputs may actually increase distrust or fear due to revealing the underlying complexities of the process in a way that is difficult - or even impossible - to grasp.

Principle of equivalence

One other challenge noted by our participants was around defining the criteria upon which an explanation should be required, and the quality that such explanation should have. One idea is to benchmark such criteria and quality to human explanations. This goes by the name of "principle of equivalence" and basically says that whenever we expect a human to explain his/her actions or decisions, we should have the same expectation for machines making decisions; and the quality and content of AI driven explanations should be the same as the one provided by humans. The principle of equivalence suggests that the same standards of disclosure for human-driven decisions should be applied to decisions that have been made or augmented by an AI system.²¹

The rich debate generated amongst the participating companies led to a more nuanced perspective. In principle, participants recognized the value of equivalency and agreed with its application when explanations are required for decisions with significant impact on people's lives. Despite this initial high level endorsement, companies avoided a binary take on the principle of equivalence, either fully supporting or totally dismissing the analogy-based argument. According to a number of our participants, that level of human-machine equivalency

In order to ensure that AI explanations are not just understandable, but also meaningful and effective, companies pointed to an audience-centric approach, providing solutions that are aligned with what the target audience is looking for. Along similar lines, one proposed solution was to enable the audience to explore explainability in a gradual manner. Bukalapak sees ways "[...] to address this issue by providing multi-level explanations. For example, start by providing an explanation at the simplest level possible [...] and then give options to the addressees of that explanation to revisit and review it at deeper levels, gaining a more detailed understanding of how AI is being used in a certain case."

depends on the specific context in which the explanation is expected, and the specific type of information that the target audience is looking for. Companies also argued that upholding a principle of equivalence could flatten the different dynamics in human-machine interaction, overlooking the strengths and limitations that each actor brings to the table.

Other participants thought that holding AI to a higher quality standard than humans by default, without looking at the specific context, may work to its detriment, and even hinder its technical advancement.

"According to current technology standards, what AI does best is improve efficiency and support decisions based on the processing of large databases; while humans can deliver higher qualified analysis within a certain amount of possible answer sets. So we should not be forcefully holding both AI and humans to the same XAI criteria and standards"

TravelFlan



elivery

Phase

Presenting & Communicating
the AI Explainability solution

In the third and final phase of the program, participating companies presented their XAI solutions by delving into user design considerations and communication strategies. This was an opportunity for companies to brief the Open Loop consortium on the progress made towards developing and delivering the XAI solutions to their intended audiences. Due to the limitations of the program,^{xxx} participants were not able to test their solutions with a representative sample of their target audiences. As an alternative, we shifted to having selected companies pitch presentations and conduct demos of their XAI solutions based on design drafts, technical documentation, mockups, and wireframes during the final program workshop. To assist the companies with their presentations, Open Loop partnered with TTC Labs and Craig Walker. TTC Labs provided participants with a

storytelling template slidedeck to help structure their presentations, along with coaching on how to deliver a pitch presentation.^{xxx} Craig Walker delivered insight talks to program participants on AI explainability and user design implications, further supporting the presentation and communication of their XAI solutions during the delivery phase. As we will see below, and even if the XAI solutions were not fully finalized or embedded into the corresponding AI applications, this alternative exercise produced interesting and enriching insights.

In this chapter we will provide a brief overview of the objectives that the companies set out to accomplish when delivering and presenting their XAI solutions, along with the policy, technical, and usability considerations involved in that process.

XAI-related objectives

Beyond and through the goal of explainability, the participants reported three additional goals that they sought to accomplish in the process of developing their XAI solutions. Firstly, the participants referenced the goal of enhancing the overall trust in AI/ML technology. The assumption here is straightforward: when users gain a better understanding of how AI-based products or services work, they will trust their outcomes. Secondly, the participants mentioned the goal of improving and refining the products or services to which the XAI solutions apply. As argued by a number of participants, explanations can create better products and services not only from a consumer / end user standpoint, but also from an AI developer perspective, as a tool for debugging code errors.^{xxxi} In fact, Nodeflux qualified XAI as an “internal problem solving tool.” And thirdly, the participating companies alluded to the goal of “getting the

record straight” regarding AI and its actual capabilities and limitations. AI explainability, as a vehicle to make AI systems understandable, has the virtue of de-mystifying the technology and of setting realistic expectations around what is actually possible and accurate to explain. In addition, setting the expectations about what AI does, or does not do in a given context, helps communicate and clarify the value proposition of the underlying AI element in the products being used or services being offered, while being clear about its potential flaws. Nodeflux, in particular, suggested also explaining ML behavior and outputs when the ML models do not work as expected. In their facial recognition XAI solution presentation, the company demo-ed a feature that would provide explainable arguments to customers when confronted with a false positive prediction or unexpected outcome.

Technical Considerations

Our participants dealt with a number of important technical considerations when tasked with building and delivering their AI explanations. Those considerations translated into a number of key procedural questions, namely how to:

- determine the sheer feasibility of explaining AI decisions and recommendations
- ensure the quality of data training sets
- implement traceability mechanisms as part of the explainability building process
- build XAI solutions that are not only cost efficient, but can easily and adequately scale

Policy Considerations

Participating companies also shared a series of policy considerations when delivering their proposed AI explanations. Three considerations are worth mentioning:

XAI's range and depth

When going through the process of presenting and delivering an explainability solution, companies often asked themselves how detailed an explanation should be and how far should companies go in opening up their books and explaining their technical modus operandi. This is related to the trade-off regarding T&E vs meaningfulness and actual understanding explained above. Nodeflux, for example, documented this challenge and - when delivering its XAI solution - posed the question of "how far should we go in terms of transparency"? Defining the right amount of information disclosed is a critical and challenging step in the pursuit of explainability goals and requirements. Apart from important elements regarding trade secrecy and incentives for innovation associated with the protection of IP rights, there are also other relevant aspects in terms of comprehensiveness and meaningfulness of the XAI solution: what to include in an explanation that reflects its complexity in an accurate manner, while still being accessible and comprehensible? This is to be decided on a case-by-case basis, but best practices should be developed to assist companies in this delicate exercise. Another related question was the extent to which users should assess metrics informing the delivery and quality of the explanation, like accuracy and precision.

The Impact of XAI on policy making

Some of the participating companies took the impact on the policy making process into account when developing their XAI solutions. Specifically, QSearch felt that the XAI should not only comply and meet the expectations of policymakers and regulators, but should also affect future policy making and regulation. In other words, the XAI solution, while complying with those expectations and (when applicable) with regulation, should also incorporate examples of practices that can then be fostered and adopted by future policy guidance and policy making processes.

The Human factor in XAI

Guided by the AI Model Governance Framework proposed matrix to help organizations determine the level of human involvement required in AI decision making,²² and its classification of the various degrees of human oversight (human-in-the loop, human-out-of-the-loop; human-over-the-loop).²³ Another important policy consideration shared by our participants revolved around how to leverage the specific role of the human in the delivery of an AI explanation. In particular, two roles were highlighted:

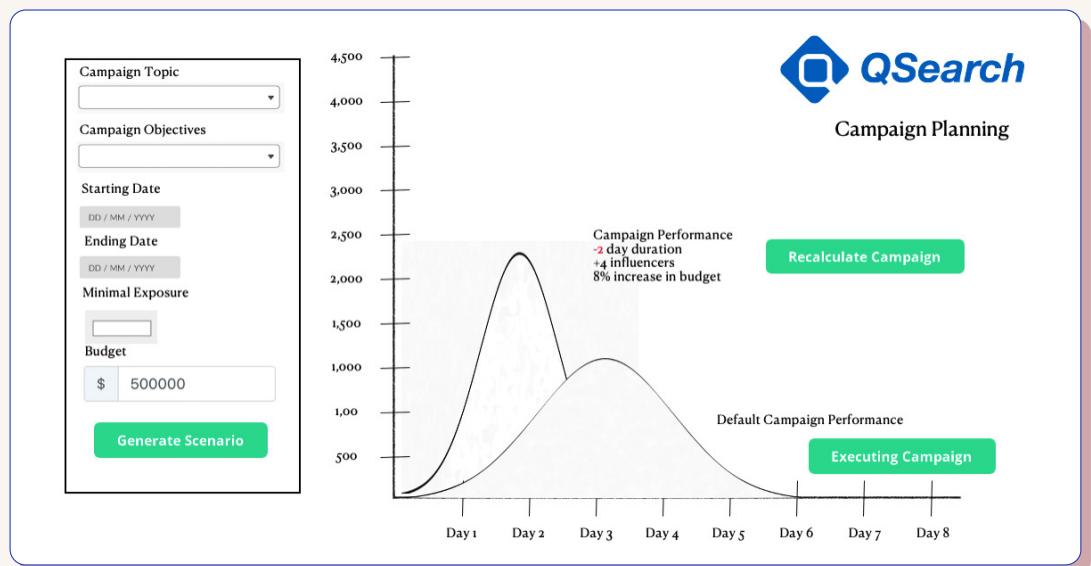
- 1 **The human as a user of the AI system**, empowered to adjust its parameters and to actively participate in the decision and recommendation made by these systems

Our findings suggest that the more opportunities there are to enable customers' actions to be part of the XAI solution, the better the user experience will be. The reasoning behind is straightforward: if you (as a user) are part of how the product works, you will be part of the explanation on how the product works.

As an example, QSearch developed a "user-controlled variable selection" feature as part of the social media campaign planning tools. The feature enables their clients to change the variables of the AI system that QSearch uses to power their services. As a company that uses AI to help its clients set up social media campaigns, find influencers to promote their products and services, and assess their impact and performance, QSearch empowered its clients to formulate different variables within the AI system to test different hypotheses regarding the selection of the influencers and the overall effectiveness of their branding campaigns. Here's how it works: once a client specifies a campaign objective, budget, and time frame, the AI system produces relevant variables and ranks each variable's likely contribution to the outcome. When low

confidence variables are identified, the client has the option of changing either the weight of the variable or changing the variable altogether. Clients will thus choose influencers from that specific time frame and for the same topic, and observe their performance as a baseline to compare with their own campaign. This feature enables clients to do a sort of A/B testing in order to generate different model outcomes and, in that way,

to project a range of different possible future outcomes. As a result, clients have the opportunity to tweak the make-up of their campaigns in order to produce a more favorable future outcome. By acting directly on the selection of the variables of the AI system and understanding their connection to the model outcome, clients not only understand how the AI system works, they become part of that process.



QSearch's Campaign Planning tool and user-controlled variable selection feature

2 The human as one who monitors and enforces the correct use of the AI system

intervening in its operation in specific cases

Besides the possibility given to users to formulate different variables for hypothesis testing, projecting different possible future outcomes, QSearch also developed a feature that would give users a remediation process to improve the AI decision. Such remediation is a way for the client to review the campaign performance, with special emphasis on the part of the campaign that did not perform as predicted. It breaks down the Influencer Marketing campaign into campaign performance and campaign

outcome, then it lists the underlying factors the system believes are correlated, and the client can adjust the factors which will generate another campaign specification to test. This feature enables human users to monitor the performance of the AI system, and to intervene in its operation by improving the way it produces decisions, predictions and recommendations.

Overall, the monitoring and intervening role of the human begs the question of the criteria determining when human intervention should be required. We found that the policy prototype provided a good foundation and starting point for companies to reflect on this question.

Usability considerations

Companies also shared a number of usability considerations when building and delivering their XAI solutions. According to the feedback received, these considerations would help enhance the customer experience when being exposed or interacting for the first time with decisions and recommendations produced by AI models (Nodeflux).

Visualization

A recurrent piece of feedback that we heard from our cohort of participants outlined the importance of using visualizations when presenting and delivering XAI solutions. The old adage “a picture is worth a thousand words” seemed to have struck a chord with the companies engaged in the program. Beyond graphical images, animations and interactive modules were also presented as key elements in the AI Explainability user design space.

Beyond the scope of this program’s list of AI applications, a good example of the power of

“Images and visual representations can better show linkages and more quickly communicate complicated relationships than words.”



visual representation and animation is the interactive tool that the Meta AI team made available to explain how Instagram Feed Ranking works, along with a slide animation that showcases the different steps the AI model goes through to order posts on a user’s feed.²⁴ Through such a tool, users can try ranking a hypothetical Instagram user’s feed, and then find out how it compares with what the feed system might predict. By going through this exercise, users gain a better understanding of how feed ranking prioritizes multiple pieces of content to help people see the posts they are most likely to find interesting, or are most likely to interact with.

How feed ranking works

Feed ranking prioritizes multiple pieces of content to help people see the posts they’re most likely to find interesting, or are most likely to interact with. The Feed ranking system predicts how likely you are to comment on it, like and save it, or tap on a profile photo. “The more likely you are to take an action, and the more heavily we weigh that action, the higher up a post will be ranked in Feed.” Check out the [Instagram blog](#) where we shed more light on this.



Now, you try it

Using the information below, try ranking a hypothetical user’s feed to see how it compares with what the feed system might predict.

Step 1: Select a profile

First, you’ll need to understand their preferences. Review the following three profiles and think about how you’d rank the posts in their feeds.

When you’re ready, select the profile you’d like to work with.



Rebekah Turner

Interests

- Fashion •••
- Cars •••
- Fitness •••

Close Friends

- Rose Padilla
- Jordan Torres



Madhu Patel

Interests

- Beauty •••
- Art •••
- Food •••

Close Friends

- Giovana Vieira
- Jordan Torres

Scores Model



Assigns a numeric value to how likely or probable you are to perform an action.

Merge Scores



The combining of various categories of scored content (i.e., posts) from images to videos, etc.

Boosts Library



A set of rules that elevates posts based on timing or user preference (i.e., holidays, close friends, etc.).

Integrity Demotions



Demotes posts that are inappropriate but do not clearly violate community guidelines.

²⁴ See: <https://ai.facebook.com/tools/system-cards/instagram-feed-ranking/>



Meta's interactive animation and tool that explains how Instagram Feed ranking works

1

First, the system gathers potential posts - excluding advertisements - from accounts you follow, like posts from friends or creators. It then removes the posts that violate our [community guidelines](#).



2

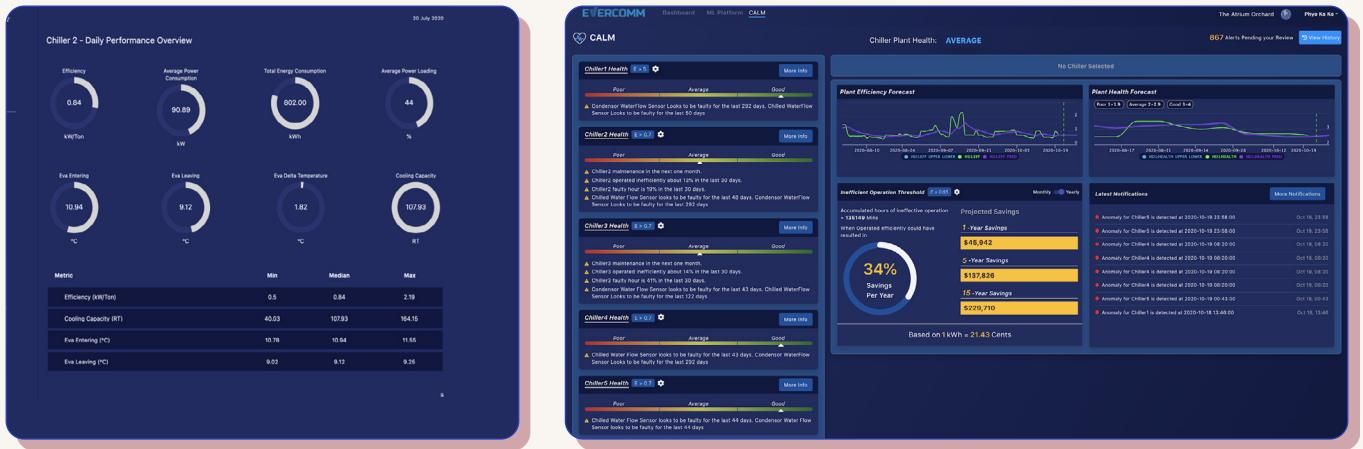
Using what's left, the system predicts how likely you are to interact with a post. To do this, it collects attributes from the post along with additional information like how often you interact with the author of a post. Based on those attributes, the model predicts how likely you are to like, save, tap or perform an action like watching a video. High model outputs (i.e., probability) indicates a higher likelihood that you're interested in the post.

Customization

Inspired by the scenario-based approach of our program, participants recommended tailoring the explanation to specific audiences. This was the case of Evercomm, a sustainability tech company that designed its XAI solution to explain how its Asset Performance Management evaluates the performance of various key equipment categories in operation (chillers, pumps, etc), and provides sustainability indicators, while tailoring it to different audiences. In its equipment performance monitoring feature,

different pieces of information can be requested by specific user groups within the same user interface, e.g. one team, focused on the assessment of environmental aspects, can access the sustainability indicators as they might need that information for a sustainability report; whereas another team can request information about how the equipment is operating via a dedicated reporting feature. The idea is to provide a specific explanation for each individual action and according to the user group performing.

In this XAI solution, different layers of information are itemized into different features, which enable different types of stakeholders within the organization to access the specific information they require in equipment performance.



Evercomm's Equipment Performance Information & Assessment.

Simplicity

As noted by several participants, XAI solutions need to be simple in order to foster meaningful understanding and increase users' adoption of the correspondent products and services. A good XAI solution should be short, simple, and clear, requiring no extra effort by the user nor additional external help. Nonetheless, we also heard from other participants the need to not oversimplify.

"An XAI solution should be explained in a way that a 5 year old would understand"



Participants referenced the delicate balance they had to make in formulating explanations that capture the complexity of the AI systems in a simple but not overly simplified manner. One way this could be done is through layered explainability approaches. The latter shows simpler language at the beginning and then adds progressively more detailed language in accordance with the user's interest in knowing more. Along similar lines, Deloitte suggested having the recipient of the explanation select the level of complexity being provided.

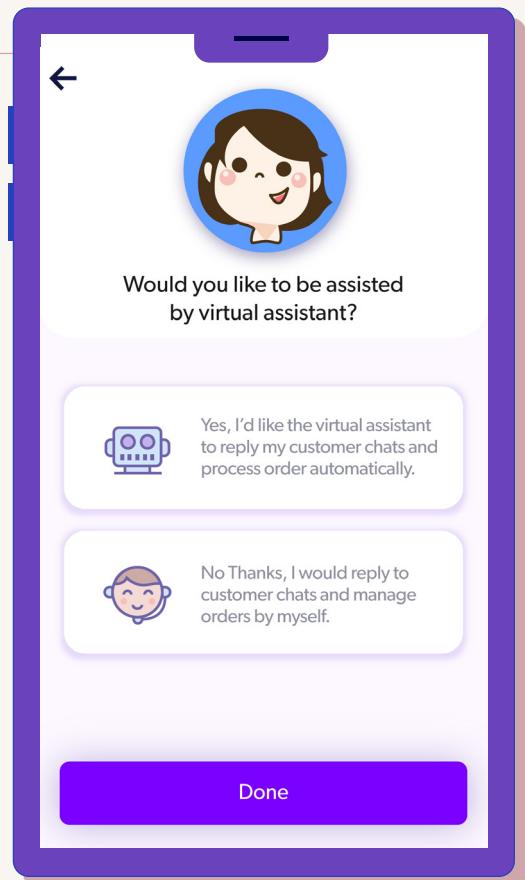
Halosis, another participating company, built a simple (yet not overly simplified) XAI solution to explain how its virtual assistant technologies

can help sellers communicate with customers and improve their sales. The new explainability feature that the company presented was developed in a way that raised the awareness of their business partners to the AI component of the feature, and to describe its use and benefits. The company developed that simple solution by a) employing the term “virtual assistant” instead of AI, presenting the latter as a helper to human action; b) by clearly differentiating the use of AI, represented by a robot icon,

Perfectly imperfect

Several of our participants stressed the importance of including in their explanations a reference to its limitations. As reminded by a number of our participants, AI never has a 100% accuracy in its predictions. This consideration debunks the myth of AI as a silver bullet for human problems and needs, and defends the value of showing AI’s limitations as a way to elicit and build trust with people that use and interact with this technology. Trabble, a self check-in system for hospitality businesses that allows guests to seamlessly check into their rooms using their own devices, enables customers to control whether they would like or not like to share sensitive information in the passport upload control feature. Specifically, the feature indicates when input data to the model cannot provide the quality or gather the conditions necessary to produce a reliable enough output. It provides a clear explanation to users about the optimal conditions for the system to work properly (disclaimer limitations), which - by explaining those limitations - can then explain possible false predictions and/or unexpected outcomes. This helps manage user expectations. Nodeflux, in its demo pitch, presented a disclaimer on AI’s limitations as part of its product offerings. “AI has limitations. We hope to make that clear in our solution, for our customer to have rational expectations on how the model behaves.” (Nodeflux)

from the use of a human counterpart; and c) by explaining the advantages of the use of AI (“reply to customer chats and process orders automatically”) versus the alternative manual / human use (“reply to customer chats and manage orders by myself”). This explanation was not only descriptive, but also actionable, empowering the user to choose between the AI and the human versions of the feature.

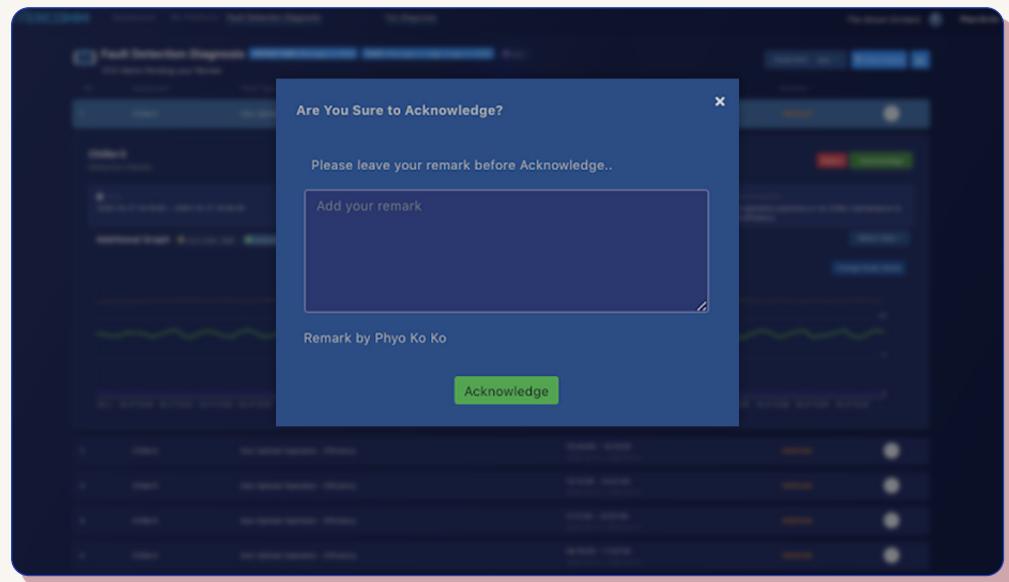


Halosis’ XAI solution on the use of virtual assistant technology

Seamless Flow

According to the feedback received, another important usability consideration is to integrate the XAI solution into the product or service in a way that flows naturally and creates a seamless ex-

perience for the user (Trabble). The explanation should not be seen as an accessory of the product; explanation should be part of the product, an almost indistinguishable part of the product.



Evercomm's Fault Detection Diagnosis.

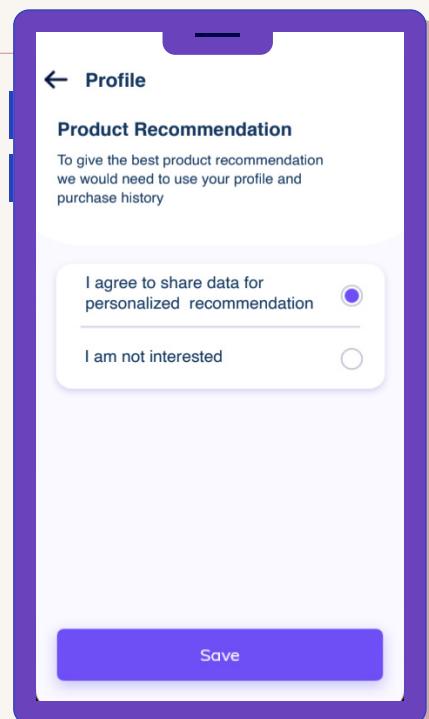
User empowerment

Another design feature for XAI solutions proposed in this phase is to empower and provide users with control options over the decision and recommendations produced by the AI/ML systems. Evercomm designed a "fault detection diagnosis" feature that collects equipment data to deliver AI generated recommendations to help governments and enterprises transition into net zero operations. This feature enables humans (namely equipment managers) to ack-

nnowledge or reject the AI driven recommendations, and to insert their feedback (input) into the ML model training process. This allows mutual interaction where the model learns from the humans and vice versa. These explanations give agency to the users and have the practical benefit of giving more data to the machine to retrain and improve the ML model.

Halosis also put user empowerment at the center of its XAI solution, by focusing on giving users the possibility to opt out. Instead of simply using their data in bulk, the company gives users the ability to turn on AI features when they are ready to share the data that is required for the particular AI feature.

Halosis' Data Usage Consent
for the use of the AI feature



Challenges from this Exercise, Learnings for Future Ones

The policy prototyping program provided valuable and rich insights. Some of these insights were gleaned directly from the challenges that we encountered in implementing the program, and the numerous lessons we learned as a result. Here are a few considerations to keep in mind for those who may want to deploy a similar approach, and for Open Loop's own forthcoming programs:

***Take into account
the technical tasks
involved in
the program and
factor these in
its overall timeline***

AI explainability was a relatively unknown field of practice across the participating companies. Despite the relatively extensive period of time allocated to this program, especially when compared to shorter, sprint-like approaches adopted elsewhere under Open Loop,²⁵ more time could have been allocated for companies to develop and implement AI explainability solutions at the technical level, that is, embedded into their product development processes and released as a feature. The time needed for that type of work may have been underestimated and, in hindsight, we may have overly relied on the technical assistance provided through our toolkit, program sessions and the mentoring sessions to accelerate those processes.

Due to time and resource constraints, participants were not able to test the explainability solutions with the audience type for which that solution was built for. To account for that, we changed and adapted the program design in a way that better met the participants' capacities. We thus converted the delivery phase into a series of demo exercises, encouraging the participating companies to advance as much

as possible their XAI solutions and still present them to an external audience, even if in a preliminary format or fashion, while acknowledging and respecting the obvious time and technical constraints. Through a dedicated hands-on working session, we helped companies build a pitch-like narrative around their XAI solutions and the status quo, including their program journey and the considerations they were building the solutions around (technical, policy and user based). A group of participants then presented their progress in the final program closing event to obtain further feedback, including from all program partners and fellow participants.

***Expect changes
in the level of
commitment of the
participating companies
and remain flexible***

We realized that the level of commitment of our participants fluctuated throughout the multi-month program. It was important to acknowledge that our participating companies were businesses whose strategic priorities might change rapidly and, as a result, whose resources could be pulled off from initially planned investments in Responsible AI practices, including the participation in a policy prototyping program. This was particularly the case given the pandemic situation in which this program took place. It is important to plan for these changes and ensure that the program can still be executed with less time and effort required from the companies, or even with a reduced number of participants. An additional number of touchpoints with the participants throughout the program, and the establishment of regular (virtual) office hours for ad hoc open discussions with the companies are a couple of ideas worth considering for future programs.

***Less is more:
simplify the program
and its requisites as
much as possible***

In this program we used a scenario-based approach. This allowed us to personalize and tailor the AI explainability tasks and experiences to each of the participating companies. Whilst this was a really effective approach, we were challenged by the sheer number of scenarios that we tried to cover in asking each company to build and follow two scenarios. While the selection of two scenarios seemed perfectly logical on paper, and was backed up by initial

desk research, designing the evaluation and testing of the policy prototype around two scenario pathways for each company was less straightforward and manageable than we had anticipated. This task ended up being overly burdensome to the companies, particularly when we inquired them about the selection of technical explainability techniques and the value-based trade-offs involved in those decisions, and when we assessed the progress made towards the delivery phase, including the building of interface solutions. For future policy prototyping programs, in order to preserve high levels of commitments and motivation from all participants, we recommend simplifying prototyping pathways and limiting them to one clear-cut scenario per participant.



5

**Policy
Insights and
Recommendations**

Recommendations for advancing AI T&E

Throughout the three phases of this Open Loop program, we collected empirical insights directly from participant companies that were actively involved in implementing XAI solutions according to our policy prototype. This resulted in a number of recommendations that we urge the policy community interested in AI T&E to consider:



Get practical

develop best practices on assessing the added value of XAI for companies and calculating its estimated implementation cost

The deployment of XAI entails significant costs for companies, from engineering expenses to compliance ones. Given the implementation costs of XAI, along with the uncertainty regarding its return on investment, there is a need for the development of best practices to assess the added value of XAI for the company and its users, along with reliable approaches to calculate the overall cost that the implementation of XAI solutions will represent to its developers.

"While the value added by XAI solution is clear and has its advantages, the overall effort and cost to add explainability to our AI product offering is unclear and uncertain"

EVERCOMM

Adopting XAI solutions can provide an added value from many perspectives. On the business side, new and compelling explainability solutions may translate into product improvement and competitive market differentiation, leading to an overall advancement in the development of more responsible AI products and services. From the regulatory viewpoint, XAI allows for greater certainty in terms of compliance, for more informed user redress mechanisms, and for higher user adoption, amongst others, which can then translate into increased user trust, additional revenues, greater brand recognition, etc.

However, as emphasized by several streams of literature, making such rapidly-evolving systems explainable, increases significantly both the amount of engineering effort and the man-

power hours for AI companies, which in turn may disadvantage smaller and less-resourced players. As a consequence, such barriers may lead to companies employing "suboptimal but easily-explained models".²⁶ The overall cost of implementation will be multifaceted, ranging from human and technical resources, to changes in engineering roadmaps and marketing strategies. AI models will also require companies to sustain the costs of maintaining and updating their AI systems, which will then entail additional costs in terms of XAI solution updates underlying those systems.

When developing their XAI regulatory guidance and/or requirements, policymakers could leverage industry and technical community's input to foster the development of these added value and implementation cost estimating practices. Through this collaborative practice, codes of practice and technical guidance could be published with specific examples of such value estimation practices and calculating approaches. This would then help the industry plan for and prepare their journey towards AI explainability, doing it so in a more confident and well-informed manner. The drafting of these technical codes and best practices would benefit from approaches typically advanced by the scientific literature, as well as regulatory bodies regarding the assessment of compliance costs. The OECD Regulatory Guidance on Compliance Cost Assessment (CCA), for instance, provides practical instructions for the calculation of regulatory costs, outlining methodologies to estimate ex ante and ex post the costs associated with adopting new provisions. In the same way, based on companies best practices and experiences, a practical guidance on the costs of XAI could be developed looking at the costs

of labor (e.g. salary costs) needed for the deployment of XAI, overhead costs, equipment, costs of external services, etc.²⁷ Relying on a clear estimation of implementation costs enhances legal certainty and allows companies to fully incorporate these expenses in their processes.



Get personal

make XAI policy guidance more personalized and context-relevant

Singapore's IMDA/PDPC have done an excellent job at capturing how its guidance on AI has been applied in practice across different sectors, stakeholders, and applications.^{xxxii} In line with this approach, we believe more work can still be done to enhance the impact and contribute to the wider adoption of the MF/ISAGO frameworks, as well as T&E frameworks in general.

One way to go about this is to further tailor parts of these frameworks to specific types of companies, stakeholders and areas of activity. Organizations developing and deploying AI systems in areas like e-commerce, pharmaceuticals, online education, employment, cybersecurity, and content moderation, will have different takes on value-based trade-offs regarding the design and deployment of their XAI solutions.

While acknowledging the impossibility - and undesirability - of covering every possible context and actor operating in this space, there is still some level of modularity that could be explored. Think, for instance, about identifying the roles within companies for which specific sections of the policy framework will be particu-

larly relevant. Or think about the specific stages of the product development processes where those sections would be most pertinent. As argued in the TTC Labs and Open Loop report on "People-centric approaches to AI Explainability", any policy solution aimed at achieving Responsible AI should factor in the contextual nature of explainability. Hence, rather than taking a one-size-fits-all approach, XAI requirements should take into account the actual need for explanations based on specific types of AI applications, their intended purpose, and the impact they have on people using or being affected by them.²⁸

These contextual and modular elements could be added to a dedicated section of AI policy frameworks, indicating - at a high level - to whom certain policy provisions are addressed. Being more explicit and granular about whom the policy is addressed to in the first place, and drafting policy guidance in a way that relates and maps to the operational day-to-day company practices, could help ensure that the policy guidance is unpacked at the right layer in the company, while increasing its overall adoption and use.



Connect the dots

create new or leverage existing toolkits, certifications and educational training modules to ensure the practical implementation of XAI policy goals

"Policies may have really good ideas on paper, but there are resources needed to make them tangible and actionable"

EVERCOMM

Several of our participants called for additional guidance that would go deeper into the practicalities of implementing XAI solutions.

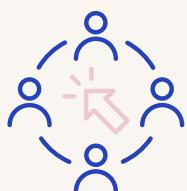
Relatedly, companies also flagged the need for more practical ways to measure the progress of their work towards the accomplishment

of the policy guidance's goals, namely through yardsticks and benchmarks;^{xxxiii} and the need for indicators of the estimated cost involved in that process. This seems to point towards the need to complement the existing policy guidance with more hands-on and experiential types of guidance and activities, containing specific parameters for XAI practical implementation, like toolkits, certificates, and educational training modules.

The good news is that there is an emerging set of tools that have been specifically created and shared for helping design and deploy AI systems in a responsible way;^{xxxiv} along with dedicated certification programs^{xxxv} and educational resources.^{xxxvi} More specifically, the OECD differentiates between three types of tools, namely technical, procedural and educational instruments that may facilitate the implementation of their AI principles for trustworthy AI. On the technical side, toolkits, software tools, technical documentation and standards can be leveraged to check for the overall reliability of AI systems. On the procedural side, guidelines providing governance frameworks, as well as risk management tools can be developed in cooperation between industry, governments and civil society organizations to document XAI procedures in a holistic and inclusive manner. Finally, when it comes to edu-

cational tools, they can play an important role in raising awareness, informing, preparing and upskilling stakeholders in the implementation of AI systems. Change management processes, capacity building tools and training programs can serve different types of audiences, from the public at large to specific affected groups.²⁹ Governments and regulators have the opportunity to "connect the dots", bridging the gap between normative guidance and practical implementation. This could be done by pointing to, adapting or integrating some of those available resources into their own guidance; or by creating new ones when the existing set of resources does not address a specific gap.

Toolkits, certifications, and educational training go deeper into what actually means to develop and deploy AI systems, including their explainability components. When these resources are connected to AI policy frameworks and regulatory guidance, companies will have a more concrete idea of the gaps they need to fill in terms of human and technical resources, as well as skills and competences. They will also gain a better understanding about the implementation challenges they will likely face and the implementation costs they will likely incur, along with a better sense of where they are in the process and what progress they have already made.



Get creative together

explore new interactive ways to co-create and disseminate policy, and increase public private collaboration

It is clear from our program that companies have a vested interest in contributing to policy making. As documented by the existing literature on XAI, the benefits generated by a multi-stakeholder approach to tackle the challenges of XAI are manifold, including more effective knowledge sharing with external actors, including end users.³⁰ Indeed, while XAI techniques are frequently employed as checks during the development process, there are still gaps when it comes to informing end users, as XAI solutions primarily serve internal stakeholders rather than external ones.³¹ Singapore has tapped into the compa-

nies' interest in co-creating shared principles and guidance for XAI by leveraging the technical input and experience of tech companies to iteratively draft and publish its AI governance framework. We very much welcome this approach and incentivize governments to follow suit and think about other innovative ways to:

- Leverage processes, tools, and practices for policy co-design and development, like citizen participation, strategic foresight, crowdsourcing, and collaborative experimentation.

- Disseminate policy findings, insights, and recommendations in more experiential formats, like use case compilations,^{xxxvii} dashboards,^{xxxviii} webinars and podcasts.^{xxxix}

The participants also encouraged regulators to gain a better understanding of the industry and follow more closely the latest scientific and technological developments. This observation stemmed from the company representatives' view that policy requires more technological knowledge in order to govern technology and better understand its industry wide implications

"[...] policy doesn't always understand its technology aspect and impact on business"



Such insufficient understanding generates not only a diffuse lack of trust in the technology itself, which is inherently complex for lay people and non tech-savvy actors, but also a more general lack of trust in AI developers.³² By collaborating with the industry in the development of

guidance and frameworks on XAI, policy makers would benefit from a deeper understanding of the inner functioning of the algorithmic systems built by companies, namely by those that are transparent about their data usage policies and the design choices made while designing and developing new products.³³ Although we cannot expect policy makers to be always up to speed on the latest advancements in emerging technologies, the establishment of discussion fora where industry players and regulators can exchange views and learn from each other will help provide policymakers with the technical knowledge they need to govern complex technologies, while bridging the trust deficit between regulators and businesses.

To increase the quality of any potential regulatory framework or policy in innovation, regulatory bodies should take a proactive approach to understanding trends in science and technology relevant for their regulatory framework; this can be done, in part, through outreach to the science and technology community through novel and forward looking policy making exercises.³⁴ We firmly believe that experimental governance projects can help support regulators in this effort.



Test and experiment

demonstrate the value and realize the potential of policy experimentation

The vast majority of the participants found that testing policy ideas around AI governance is a relevant endeavor as it can help policymakers understand the challenges that companies may encounter when asked to follow its guidance in terms of technical feasibility and business viability. Regarding AI Explainability in particular, companies confirmed that:

"it is important to test it [XAI policy] at a smaller scale before it's fully deployed"



Moreover, experimental approaches, such as policy prototyping, can help build effective policies that also allow tech businesses to better absorb and integrate normative provisions in product development stages. As shown in previous policy prototyping experiences, a step-by-step controlled testing on a smaller scale would significantly benefit companies, particularly less-resourced ones.³⁵

Singapore IMDA/PDPC has actively engaged in the implementation of regulatory sandboxes for privacy when updating its Personal Data Protection Act and drafting its Advisory Guidelines on the Enhanced Consent Framework,³⁶

and has also helped this particular policy prototyping program on AI Explainability with Open Loop to inform its ongoing work on AI Governance. These are two examples of government-industry partnerships in experimental governance endeavors. More examples are needed, and an additional and more diversified set of actors from academia, civil society,

and technical communities is warranted. Sandboxes and prototyping programs have the potential to shape policy and inform future laws and other governance instruments in a truly evidence-based way, but for that to happen we need to deploy them more frequently, and assess their impact more consistently. Open Loop is a step in that direction.



6



onclusion

Despite all the technical, policy, and regulatory advancements in the field of AI and Responsible AI, these are still early days for AI explainability. In fact, companies are just starting to conceptualize and put into practice XAI solutions. One of the main merits of this Open Loop program, as documented by the participants, was not only to introduce and raise awareness to the topic of XAI, but also to equip them with the know-how, tools, and techniques to build and implement XAI solutions in practice, and in a responsible way.

"The biggest learning for the program for us was the ability to raise awareness for XAI in our company"

nodeflux

The program, described by some of the companies as "eye opening", played an important role in consolidating XAI as a key element of the AI product development processes within their companies.

"Raising internal awareness is the first step on our XAI journey. In order to implement XAI for our users, our organization needs to first assimilate the benefits of this effort for ourselves. Once we do that then more ideas and approaches to making XAI a living reality will come within the organization"

 Halosis

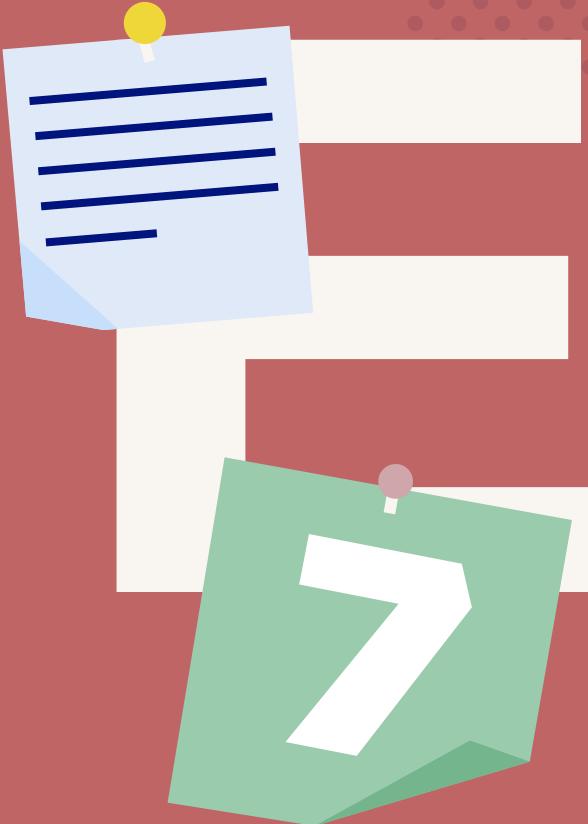
This Open Loop initiative produced rich insights on how to operationalize AI explainability, and - hopefully - paved the way for further experimentation in this field. In fact, there is still research to be done in order to further advance the field of XAI and contribute to its wider adoption and practical implementation. The role of XAI in building higher levels of trust with users, and improving the quality of the corresponding product or service, is an important assertion that we captured through our participants' feedback on the program: XAI "will directly change how our users see and use our products, elevating their expectations regarding the quality of our products" (Travelflan). Further research to confirm, demonstrate, and elaborate on this type of finding would be of most value in enhancing the value and need of AI explainability.

Another important element of the program was the community sentiment and the collaborative modus operandi that was generated amongst the participants. Companies stated how beneficial it was to exchange information, knowledge, and practices with other companies, learning from each other throughout the program. The participants also highlighted how that community spirit challenged them to excel in the program and leverage its resources in the best possible way to build and implement XAI solutions.

With Open Loop's experimental and multi-stakeholder, consortium-driven approach, we hope to continue broadening the perspectives involved in the responsible AI - and wider AI governance - debate by enriching it with input grounded in qualitative and community-generated evidence. And, in this process, we encourage policymakers to join our efforts and embark on similar experimental governance programs.



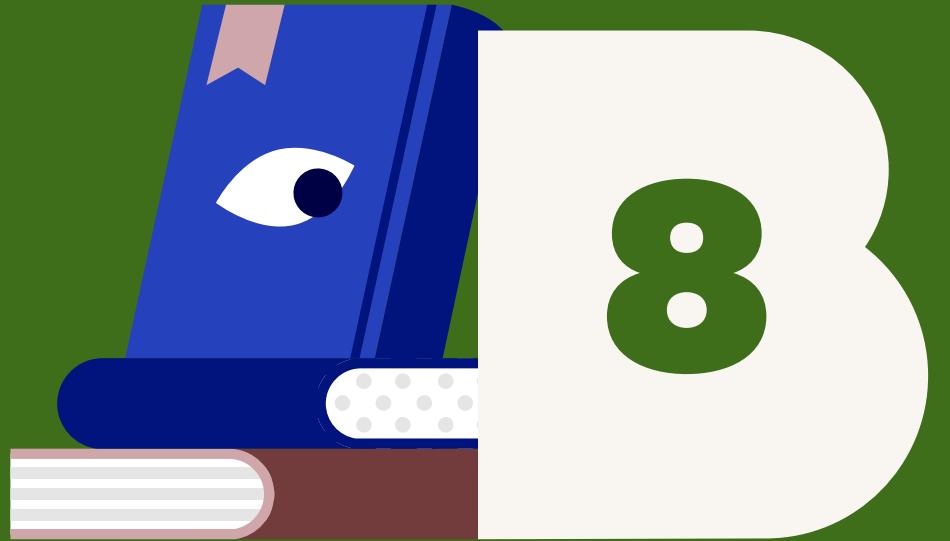
ndnotes



- I Now part of Freed Group.
- II Previously Facebook, see <https://about.fb.com/news/2021/10/facebook-company-is-now-meta/>
- III The policy prototype we used for testing and evaluation of Singapore's MF and ISAGO frameworks consisted of a curated selection of relevant provisions on T&E, mostly taken from MF's sections Foreword, Introduction (Objectives, Guiding principles for the Model Framework), and Model AI Governance Framework (Operations Management, Stakeholder Interaction and Communication); and ISAGO's section Operations Management and Stakeholder Interaction and Communication. The policy prototype can be found in the annex section of this report.
- IV We follow the definition of AI system proposed by the OECD: "machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments." An AI system may comprise several ML models. See OECD 2021a.
- V This involved supplementary explanation strategies and tools to uncover the feature importance (and interactions), generate a simpler model, or provide context through counterfactuals.
- VI The AI Transparency and Explainability Technical Guidance can be found in the annex section of this report.
- VII Throughout this report, we will use the acronyms "T&E" to refer to Transparency and Explainability as a whole, and "XAI" to refer to AI explainability specifically. "Explainable AI", "AI Explainability" and "XAI" are used interchangeably in this report.
- VIII As noted by the OECD, the output of an AI system may consist of predictions, recommendations or decisions. See OECD 2021a.
- IX ML is a sub-discipline to AI but, for the sake of simplicity, we will be using these terms interchangeably in this report.
- X The second edition of the Model Framework, which included the Implementation and Self Assessment Guide for Organisations (ISAGO), was launched at the World Economic Forum Annual Meeting in Davos, Switzerland in January 2020. For more information, see <https://oecd.ai/fr/wonk/singapores-model-framework-to-balance-innovation-and-trust-in-ai> The ISAGO can be found at <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf>
- XI The policy prototype, which included selected provisions on T&E from Singapore's MF and ISAGO frameworks, can be found in the annex section of this report.
- XII See the section on "Methodology" in this report for more details.
- XIII For more information, see <https://www.imda.gov.sg/Who-We-Are/about-imda>
- XIV For more information, see <https://aisingapore.org/>
- XV The original partnership collaboration was established between BasisAI and Open Loop. BasisAI was acquired by Temasek-founded Aicadium in August 2021. For more information, see <https://aicadium.ai/>
- XVI For more information, see <https://www.ttclabs.net/>
- XVII For more information, see <http://craigwalker.com.au/>
- XVIII Inspiration for the definition of the scenario categories stemmed from Singapore's Model Framework and from the latest developments in XAI literature and regulatory guidance, including the Information Commissioner's Office (ICO) and Alan Turing Institute Explainability project. See ICO 2018.
- XIX Given that each company selected two scenarios, our scenario-based methodology encompassed 24 possible scenarios through which to evaluate and test the Singapore's governance frameworks' provisions included in our AI T&E prototype.

- XX The AI Transparency and Explainability Technical Guidance can be found in the annex section of this report.
- XXI For an overview of mobile ethnography platforms and more information about the platform we worked with in this program see, e.g. <https://www.insightplatforms.com/10-platforms-for-mobile-ethnography/> and <https://dscout.com/>, respectively.
- XXII This was the case of the so-called: “principle of equivalence” (e.g. section 3.48 in the MF, p. 53), according to which the same standards of disclosure for human-driven decisions should be applied to decisions that have been made or augmented by an AI system.
- XXIII All quotes in this report are statements shared by the participating companies throughout our program. They were collected upon instances of testing the policy prototype, mainly through the mobile ethnography application and the program workshops. See also the section on “Mobile Ethnography” in this report for more details.
- XXIV Participants were explicitly asked to not disclose any proprietary, confidential information, but to explain in their own words and at a fair level of abstraction the technical elements of the algorithms and models they were developing without getting into overly specific details.
- XXV The platform used by the participants was Bedrock, a machine learning operations (MLOps) platform made available by Aicadium. It enables rapid and responsible deployment of machine learning algorithms into production. It allows AI developers to peer inside the “black box” of AI systems within their organization, and achieve explainability, maintainability and auditability in-built into their AI system. Find a Bedrock intro video on YouTube <https://youtu.be/mrjD0mdviXg> or additional information at <https://www.techinasia.com/temasekbacked-startup-aims-open-ai-black-box>
- XXVI As referenced previously, the AI Transparency and Explainability Technical Guidance can be found in the annex section of this report.
- XXVII Note that we did not test this guidebook or ask for express feedback on it. For insights from testing policy text and guides to implementation please see other Open Loop programs, e.g., the Open Loop Europe on Automated Decision-making Impact Assessments or Open Loop Mexico on AI Transparency and Explainability (report on the latter forthcoming in 2022).
- XXVIII The trade-off between accuracy and interpretability has been contested in technical studies. In this regard, experts have argued that, for high stakes decisions, the way forward should be to design models that are inherently interpretable, instead of trying to explain black box models. See Rudin 2019.
- XXIX See the section “Challenges from this exercise, learnings for future ones” in this report.
- XXX The storytelling template and proposed narrative contained the following: a recap of who the company is and what it does; main company audience; main challenges related to building XAI; an overview of the policy, technical, and usability considerations made by the companies in building XAI solutions throughout the program.
- XXXI This is in line with our XAI literature review. See the: “Introduction” section in this report for more details. See, for example, “Compendium of use cases: practical illustrations of the Model AI Governance Framework.” [Singapore IMDA / PDPC](#)
- XXXII See, for example, “Compendium of use cases: practical illustrations of the Model AI Governance Framework.” [Singapore IMDA / PDPC](#)
- XXXIII See the section on “Policy Effectiveness” in this report.
- XXXIV See OECD’s Digital Economy Paper on “Tools for trustworthy AI: a framework to compare implementation tools for trustworthy AI systems.” See OECD 2021a. The framework aims to help collect, structure and share information, knowledge and lessons learned on tools, practices and approaches for implementing trustworthy AI. See also TechTransformed for “a set of practical resources to help organizations turn big dilemmas to hands-on decisions” at <https://doteveryone.org.uk/techtransformed/>

- XXXV See, for example, the Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), created by the Institute of Electrical and Electronics Engineers (IEEE) with the goal of advancing specification and making processes that advance transparency, accountability and reduction in algorithmic bias in autonomous and intelligent systems. In Singapore, the School of Computer Science and Engineering at Nanyang Technological University, together with the Singapore Computer Society, offers a Certificate in AI Ethics and Governance, which contains a module on "Governance for AI Explainability".
- XXXVI See, for example, "Elements of AI", a series of free online courses on the basics of AI, created by Reaktor and the University of Helsinki.
- XXXVIII See, for example, "Compendium of use cases: practical illustrations of the Model AI Governance Framework". [OECD AI Policy Observatory's live repository of over 700 AI policy initiatives across the world.](#)
- XXXIX See for example, UK's Information Commissioner Office (ICO) hosted webinars and podcasts recordings. <https://ico.org.uk/for-organisations/webinars-and-podcasts/>



ibliography

Accenture. "Responsible AI From principles to practice." (2021) https://www.accenture.com/_acnmedia/PDF-149/Accenture-Responsible-AI-Final.pdf#zoom=50

Andrade, Norberto Nuno Gomes, and Verena Kortscheder. "AI Impact Assessment: A Policy Prototyping Experiment." (2021) https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf

Ananny, Mike, and Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *new media & society* 20, no. 3 (2018): 973-989. http://ananny.org/papers/anannyCrawford_seeingWithoutKnowing_2016.pdf

Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. "Explainable machine learning in deployment." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648-657. (2020a). <https://dl.acm.org/doi/abs/10.1145/3351095.3375624>

Bhatt, Umang, McKane Andrus, Adrian Weller, and Alice Xiang. "Machine learning explainability for external stakeholders." (2020b) [arXiv preprint arXiv:2007.05408 \(2020b\)](https://arxiv.org/pdf/2007.05408.pdf). <https://arxiv.org/pdf/2007.05408.pdf>

Blind, Knut. "The impact of regulation on innovation." *NESTA Working Paper Series*, Paper No. 12/02 (2021) <https://www.nesta.org.uk/report/the-impact-of-regulation-on-innovation/>

Business at OECD (BIAC). "Regulatory Sandboxes for Privacy Analytical Report." November. (2020). <https://biac.org/wp-content/uploads/2021/02/Final-Business-at-OECD-Analytical-Paper-Regulatory-Sandboxes-for-Privacy.pdf>

Dimanov, Botty, Umang Bhatt, Mateja Jamnik, and Adrian Weller. "You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods." In *SafeAI@ AAAI*. (2020). <http://ceur-ws.org/Vol-2560/paper8.pdf>

Doshi-Velez, Finale, and Mason Kortz. "Accountability of AI Under the Law: The Role of Explanation." Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, (2017). <http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584>

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51, no. 5 (2018): 1-42. <https://dl.acm.org/doi/pdf/10.1145/3236009>

Gunning, David, Mark Stefk, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. "XAI - Explainable artificial intelligence." *Science Robotics* 4, no. 37 (2019). <https://www.science.org/doi/full/10.1126/scirobotics.aay7120>

ICO, Alan Turing Institute. "Project explAIin." Interim Report. (2018) <https://ico.org.uk/media/2615039/project-explain-20190603.pdf>

IMDA and PDPC. "Model Artificial Intelligence Governance Framework." (2020). <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf>

Lucic, Ana, Madhulika Srikumar, Umang Bhatt, Alice Xiang, Ankur Taly, Q. Vera Liao, and Maarten de Rijke. "A Multistakeholder Approach Towards Evaluating AI Transparency Mechanisms." (2021). <https://arxiv.org/abs/2103.14976>

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems." *ACM Transactions on Interactive Intelligent Systems (TiIS)* 11, no. 3-4 (2021): 1-45. <https://arxiv.org/pdf/1811.11839.pdf>

OECD. "Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449." (2021a). <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

OECD(b). "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems", OECD Digital Economy Papers, No. 312, OECD Publishing, Paris. (2021b). <https://doi.org/10.1787/008232ec-en>

OECD. "OECD Regulatory Compliance Cost Assessment Guidance". OECD Publishing. Paris (2014)

Rossi, Francesca. "Building trust in artificial intelligence." Journal of international affairs 72, no. 1 (2018): 127-134. <https://www.jstor.org/stable/26588348>

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215. <https://arxiv.org/pdf/1811.10154.pdf>

Selbst, Andrew D., and Solon Barocas. "The intuitive appeal of explainable machines." Fordham L. Rev. 87 (2018): 1085. <https://par.nsf.gov/servlets/purl/10121294>

TTC Labs, Report on "People-Centric Approaches to AI Explainability: Insights from product and policy prototyping with startups." (2022). https://downloads.ctfassets.net/94xygyiuusop/4Eu9tBbxKaNfsvKSMfOOWe917e9d900497423c08419916d5cc284/TTC_Labs_Approaches_to_AI_Explainability.pdf

Verma, Sahil, John Dickerson, and Keegan Hines. "Counterfactual explanations for machine learning: A review." arXiv preprint arXiv:2010.10596 (2020). <https://arxiv.org/abs/2010.10596>

Vilone, Giulia, and Luca Longo. "Explainable artificial intelligence: a systematic review." arXiv preprint arXiv:2006.00093 (2020). <https://arxiv.org/pdf/2006.00093.pdf>

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841. <https://arxiv.org/pdf/1711.00399.pdf>

Wang, Yichuan, Mengran Xiong, and Hossein Olya. "Toward an understanding of responsible artificial intelligence practices." In Proceedings of the 53rd hawaii international conference on system sciences, pp. 4962-4971. Hawaii International Conference on System Sciences (HICSS), (2020). <https://eprints.whiterose.ac.uk/162719/8/Toward%20an%20Understanding%20of%20Responsible%20Artificial%20Intelligence%20Practices.pdf>

World Economic Forum (WEF) and IMDA. "Implementation and Self-Assessment Guide for Organisations". ISAGO. (2020) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf>



nnexes

AI Transparency and Explainability Prototype

Model AI Governance Framework - Second Edition	73
Foreword.....	73
Introduction	73
Model AI Governance Framework	74
Operations Management.....	74
Algorithm and Model - Explainability	74
Stakeholder Interaction and Communication	75
General disclosure.....	75
Policy for explanation	76
Bringing explainability and transparency together in a meaningful way	76
Companion to the Model AI Governance Framework	
- Implementation and Self-Assessment Guide for Organizations (ISAGO)	77
Operations Management.....	77
Stakeholder Interaction and Communication.....	78

The policy prototype we used for testing and evaluation of Singapore's Model AI Governance Framework (MF), and its Implementation and Self-Assessment Guide for Organizations (ISAGO), consisted of a curated selection of their most relevant provisions on transparency and explainability. The policy prototype consists of excerpts from MF's Foreword, Introduction (Objectives, Guiding principles for the Model Framework), and Model AI Governance Framework (Operations Management, Stakeholder Interaction and Communication) sections; as well as from ISAGO's Operations Management and Stakeholder Interaction and Communication sections.

Model AI Governance Framework - Second Edition

Foreword

In January 2019, Singapore launched our Model AI Governance Framework (Model Framework) at the World Economic Forum in Davos. The Model Framework's unique contribution to the global discourse on AI ethics lies in translating ethical principles into practical recommendations that organizations could readily adopt to deploy AI responsibly.

The ISAGO complements the Model Framework by allowing organizations to assess the alignment of their AI governance practices with the Model Framework, while providing useful industry examples and practices.

These initiatives play a critical role in Singapore's National AI Strategy. They epitomize our plans to develop a human-centric approach towards AI governance that builds and sustains public trust. They also reflect our emphasis on co-creating an AI ecosystem in a collaborative and inclusive manner. The Model Framework and ISAGO will pave the way for future developments, such as the training of professionals on ethical AI deployment, and laying the groundwork for Singapore, and the world, to better address AI's impact on society.

(p.7 and 8)

Introduction

Objectives

The exponential growth in data and computing power has fuelled the advancement of data-driven technologies such as AI. AI can be used by organizations to provide new goods and services, boost productivity, enhance competitiveness, ultimately leading to economic growth and a better quality of life. As with any new technology, however, AI also introduces new ethical, legal and governance challenges. These include risks of unintended discrimination potentially leading to unfair outcomes, as well as issues relating to consumers' knowledge about how AI is involved in making significant or sensitive decisions about them.

(Section 2.1, p.13)

The extent to which organizations adopt the recommendations in this Model Framework depends on several factors, including the nature and complexity of the AI used by organizations, the extent to which AI is employed in the organizations' decision-making, and the severity and probability of the impact of the autonomous decision on individuals.

(Section 2.5, p.14)

Guiding principles

The Model Framework is based on two high-level guiding principles that promote trust in AI and understanding of the use of AI technologies:

- Organizations using AI in decision-making should ensure that the decision-making process is **explainable, transparent** and **fair**.

Although perfect explainability, transparency and fairness are impossible to attain, organizations should strive to ensure that their use or application of AI is undertaken in a manner that reflects the objectives of these principles as far as possible. This helps build trust and confidence in AI.
- AI solutions should be **human-centric**.

As AI is used to amplify human capabilities, the protection of the interests of human beings, including their **well-being** and **safety**, should be the primary considerations in the design, development and deployment of AI.

(Section 2.7, p.15)

Model AI Governance Framework - Operations Management

Algorithm and Model - Explainability

Explainability is achieved by explaining how deployed AI models' algorithms function and/or how the decision-making process incorporates model predictions. The purpose of being able to explain predictions made by AI is to build understanding and trust. An algorithm deployed in an AI solution is said to be explainable if how it functions and how it arrives at a particular prediction can be explained. When an algorithm cannot be explained, understanding and trust can still be built by explaining how predictions play a role in the decision-making process.

(Section 3.26, p.44)

Organizations deploying AI solutions are recommended to adopt the following practices:

- Model training and selection are necessary for developing an intelligent system (i.e. a system that contains AI technologies). Documenting how the model training and selection processes are conducted, the reasons for which decisions are made, and measures taken to address identified risks will enable the organization to provide an account of the decisions subsequently.

In this regard, the field of Automated Machine Learning aims to automate a significant portion of machine learning workflows, including feature engineering, feature selection, model selection and hyper-parameter tuning. Organizations using these types of tools can consider the transparency, explainability and traceability of the automated machine learning approach, as well as the models selected.

- Incorporating descriptions of the solutions' design and expected behavior into product or service descriptions and system technical specifications documentation demonstrates accountability to individuals and/or regulators. This could also include design decisions in relation to why certain features, attributes or models are selected in place of others. These steps can help provide greater clarity on an AI model by giving understandable and digestible insights into how the model operates.

Where an organization's AI system was obtained or procured from a third-party AI solution provider, the organization can consider requesting assistance from the AI solution provider as they may be better placed to explain how the solution functions.

- Supplementary explanation tools are helpful for explaining AI models, especially models that are less interpretable (also known as "black box" systems). These tools help make the underlying rationale of an AI system's output more interpretable and intelligible to those who use the system. It is possible to use a combination of these tools to improve the explainability of an AI model's decision.

[These tools are known as "supplementary" as there is at present no single comprehensive technical solution for making AI models explainable. These tools thus play a supplementary role in providing some level of interpretability on an AI model's operation. Examples of these tools include the use of surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or Self-Explaining and Attention-Based Systems.]

(Section 3.27, p.44 and 45, incl. footnote)

Technical explainability may not always be enlightening, especially to the man on the street. Implicit explanations of how the AI models' algorithms function may be more useful than explicit descriptions of the models' logic. For example, providing an individual with counterfactuals (such as "you would have been approved if your average debt was 15% lower") and/or comparisons (such as "these are users with similar profiles to yours that received a similar decision") can be a powerful type of explanation that organizations could consider.

(Section 3.28, p.45)

Stakeholder Interaction and Communication

This section is intended to help organizations take appropriate steps to build trust in the stakeholder relationship strategies when deploying AI.

General disclosure

Organizations are encouraged to provide general information on whether AI is used in their products and/or services. Where appropriate, this could include information on what AI is, how AI is used in decision-making in relation to consumers, what are its benefits, why your organization has decided to use AI, how your organization has taken steps to mitigate risks, and the role and extent that AI plays in the decision-making process. For example, an online portal may inform its users that they are interacting with an AI-powered chatbot and not a human customer service agent.

Organizations can consider disclosing the manner in which an AI decision may affect an individual consumer, and whether the decision is reversible. For example, an organization may inform the individuals that their credit ratings may lead to a loan refusal not only from this organization but also from other similar organizations, while also informing them that such a decision is reversible if individuals can provide more evidence on their credit worthiness.

Policy for explanation

Organizations are encouraged to develop a policy on what explanations to provide to individuals and when to provide them. Such policies help ensure consistency in communication, and clearly sets out roles and responsibilities of different members of your organization. These can include explanations on how AI works in an AI-augmented decision-making process, how a specific decision was made and the reasons behind that decision, and the impact and consequence of the decision. The explanation can be provided as part of general communication. It can also be information in respect of a specific decision upon request. In this regard, the principle of equivalence can provide some guidance such that the same standards of disclosure for human-driven decisions is applied to decisions that have been made or augmented by an AI system.

Bringing explainability and transparency together in a meaningful way

Appropriate interaction and communication inspire trust and confidence as they build and maintain open relationships between organizations and individuals (including employees). Stakeholder relationship strategies should also not remain static. Companies are encouraged to test, evaluate and review their strategies for effectiveness. Further, the extent and mode of implementation of these factors could vary from scenario to scenario.

As different stakeholders have different information needs, an organization can start by first identifying its audience (i.e. its external and internal stakeholders). An organization's external stakeholders may include consumers, regulators, other organizations it does business with, and society at large. Its internal stakeholders may include the organization's board, management and employees. An organization can also consider the **purpose** and the **context** of the interaction with its stakeholders. For the purposes of illustration, this Model Framework provides considerations for interacting with consumers and other organizations.

(Sections 3.45 to 3.50, p.53 and 54)

Companion to the Model AI Governance Framework

– Implementation and Self-Assessment Guide for Organizations (ISAGO)

Operations Management

Algorithm and Model - Explainability

Can your organization explain how the deployed AI model functions and arrives at a particular prediction?

To enhance explainability, consider:

- Implementing supplementary explanation strategies to explain AI models, especially for models that are less interpretable. Examples of these strategies include the use of surrogate models, partial dependence plots, global variable importance/interaction, sensitivity analysis, counterfactual explanations, or self-explaining and attention-based systems. These strategies help make the underlying rationale of an AI system's output more interpretable and intelligible to those who use the system. It is possible to use a combination of these strategies to improve the explainability of an AI model's decision
- Generating model reports that contain the level of explainability of each feature
- Putting in place a factsheet outlining the details on how the AI model operates, including how the model was trained and tested (with what types of data), its performance metrics, fairness and robustness checks, intended uses and maintenance
- Developing a forecasting model that mimics the dynamics of the real-world business situation that is in line with the user's expectation of the business dynamics
- Training a simpler version of the model to provide better explanation about the inner workings of the complex model
- Having assessed trade-offs, use simpler models such as linear regression instead of more complex ones like neural networks
- Identifying and explaining model limitations to minimize potential for misuse

Consider whether it is relevant to request assistance from the AI solution provider to explain how the identified AI solution functions

Consider whether it is useful to use visualizations (e.g. graphs) to explain technical predictions at the model and the individual level

Consider whether it is useful to explain decisions in narrative terms (e.g. correlation between factors) and use simple indicators to measure output/ outcomes (e.g. use "high/medium/low" instead of percentages to measure risk aversion)

Consider documenting information/guiding descriptors (e.g. database description, model description, evaluation parameters) for AI modelling outputs to provide insights on major contributing factors of each model

Consider using the Local Interpretable Model-Agnostic Explanations (LIME) technique to explain contributing factors that drive the output of the AI model and SHapley Additive exPlanation (SHAP) to explain how much a particular feature contributed to the decision of the AI model, and related techniques (e.g. Leave One Covariate Out, or LOCO, counterfactual, partial dependence and Individual Conditional Expectation, or ICE to explain the importance of a feature and how the values of that feature affect the outcome

(Section 4.24, p.23)

Stakeholder Interaction and Communication

Operationalizing communication strategy based on purpose and audience

Has your organization identified the various internal and external stakeholders that will be involved and/or impacted by the deployment of the AI solution?

Did your organization consider the purpose and the context under which the explanation is needed?

Did your organization tailor the communication strategy and/or explanation accordingly after considering the audience, purpose and context?

Where practical and/or relevant, consider:

- Customizing the communication message for the different stakeholders who are impacted by the AI solution
- Providing different levels of explanation at:
 - Data (e.g. types and range of data used in training the algorithm)
 - Model (e.g. features and variables used and weights)
 - Human element (e.g. nature of human involvement when deploying the AI system)
 - Inferences (e.g. predictions made by the algorithm)
 - Algorithmic presence (e.g. if and when an algorithm is used)
 - Impact (e.g. how the AI solution affects users)
- After identifying the audience, purpose and context, organizations should consider prioritizing what needs to be explained to the different stakeholders
- Providing process-based explanation (e.g. considerations on the data used, model selection and steps to mitigate risk of the AI solution) and/or outcome-based explanation (i.e. the purpose and impact/consequences of the AI solution on users)
- Both the language and complexity of concepts in communication, and use heuristics for stakeholders that are less technical

- Consider charting the stakeholder journey and identifying the type of information, level of details and objective of informing the customer at each significant milestone. This could minimize information fatigue

Did your organization inform relevant stakeholders that AI is used in your products and/or services?

In disclosing information to relevant stakeholders, consider:

- Disclosing to consumers which data fields were most important to the decision-making process and the values in those data fields
- Whether it is relevant to provide information at an appropriate juncture on what AI is and when, why and how AI has been used in decision-making about the users. Organizations could also document and explain the reason for using AI, how the AI model training and selection processes were conducted, the reasons for which decisions were made, as well as steps to mitigate risks of the AI solution on users. By having a clear understanding of the possible consequences of the AI-augmented decision-making, users could be better placed to decide whether to be involved in the process and anticipate how the outcomes of the decision may affect them
- Whether it is necessary to provide information on the role and extent that AI played in the decision-making process (e.g. statistical results and inferences) in plain language and in a way that is meaningful to the individuals impacted by the AI solution (e.g. infographics, summary tables and simple videos). Organizations could also use decision trees or simple proxy model representations to visualize complexity and justify decisions by the AI model to stakeholders

(Sections 5.1 to 5.4, p.29 and 30)

Note: Further sections and policy provisions on AI Transparency & Explainability from Singapore's MF and ISAGO were tested through 1:1 interviews. Those sections and provisions were selected based on the scenario components that the participating companies had selected in the beginning of the program (audience, context, purpose, content), and on the AI Explainability solutions that they chose to build accordingly.

AI Transparency & Explainability Technical Guidance

Playbook for AI transparency & Explainability	82
Step 1: Determine the risk posed by the AI system and its outputs	82
Step 2: Determine specific requirements	82
Step 3: Assess whether to use globally or locally explainable models	82
Step 4: Determine the goals for understanding and the associated target audience	83
Step 5: Determine the appropriate disclosure per target audience	83
Step 6: Collect the relevant elements of your explanation.....	83
Step 7: Implement the technical and organizational measures needed for the explanation	83
Step 8: Assess and evaluate	83
Overview of direct explainable algorithms	84
Linear regression (LR).....	84
Logistic regression (LR)	85
Regularized regression (LASSO and Ridge).....	86
Generalized linear model (GLM)	87
Generalized additive model (GAM).....	88
Decision tree (DT)	88
Rule/decision lists and sets	90
Supersparse linear integer model (SLIM)	91
Naive Bayes	92
K-nearest neighbor (KNN)	92
Overview of indirectly interpretable algorithms	93
Support vector machines (SVM).....	93
Artificial neural net (ANN)	94
Random Forest	95
Ensemble methods	96
Supplementary explanation strategies and tools	96
Surrogate models (SM) [Post-hoc, local & global]	97
Partial Dependence Plot (PDP) [Post-hoc global]	98
Individual Conditional Expectations Plot (ICE) [Post-hoc local].....	99
Accumulated Local Effects Plots (ALE) [Post-hoc global]	100
Global Variable Importance [Post-hoc global].....	101
Global Variable Interaction [Post-hoc global].....	102
Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP) [Post-hoc local (possibly global)].....	104

Local Interpretable Model-Agnostic Explanation (LIME) and anchors [Post-hoc local]	104
Shapley Additive ExPlanations (SHAP) [local post hoc]	106
Counterfactual Explanation [Post-hoc local].....	107
Self-Explaining and Attention-Based Systems.....	108
Explanation toolkits and frameworks	109
Tools for extracting model explanations	110
DrWhy.AI (including DALEX)	112
alibi	113
skater.....	114
Tf-explain.....	114
iml	115
iNNvestigate.....	115
treeinterpreter.....	116
Captum.....	116
Causalml	116
DeepExplain	117
grad-cam.....	117
Keras-vis	118
Interpret-ml.....	118
Tools for detecting Bias	119
AI Fairness 360	119
MI Fairnes-gym.....	119
Examples UX design	120
Projects by If	120
TTC Labs.....	120

Glossary of terms	121
-------------------	-----

This document provides an overview of the current AI explainability field, from a technical perspective. In particular, it provides an overview of white-box algorithms, and black-box algorithms that need supplementary strategies to be explainable.

For each strategy we provide background information, articles, tutorials, R packages, and Python libraries

We conclude with an overview of the most prominent toolkits that combine several strategies and support for various architectures.

This document is a starting point for explaining your algorithms, directing you to resources that can help you further. It does not serve as a complete introduction into the technical aspects of AI explainability. This document incorporates elements from the following resources:

- Information Commissioners Office (2020). Explaining decisions made with AI: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/>
- Molnar (2020). Interpretable Machine Learning: <https://christophm.github.io/interpretable-ml-book/>
- James et al (2017). An Introduction to Statistical Learning <http://faculty.marshall.usc.edu/gareth-james/ISLR/ISLR%20Seventh%20Printing.pdf>

1.1. Playbook for AI transparency & Explainability

Explaining AI decisions might seem as a technical issue, however the explanation and the delivery require a broader perspective as shown in this playbook.



Determine the risk posed by the AI system and its outputs

Based on the outcomes of your risk assessment, determine what risks the AI system and its outputs pose for individual and/or collective values.



Determine specific requirements

Based on this assessment determine what level of information, disclosure and interpretability is required.



Assess whether to use globally or locally explainable models

In some cases, consider using globally explainable models instead of models that can only locally (on the instance level) be explained (see Overview of direct explainable algorithms).



**STEP
4**

Determine the goals for understanding and the associated target audience¹

Determine what the goal of information provision and disclosure is. Based on the different goals, determine which target audience must be addressed



**STEP
5**

Determine the appropriate disclosure per target audience

For each target audience determine what form of disclosure or explainability is required and desirable. When it comes to interpretability and explainability, tailor the explanation and message to the goals of the target audience and their ability to understand and assess the information that you provide them. Take into account factors such as expertise and time limitation.



**STEP
6**

Collect the relevant elements of your explanation

There are several aspects of the algorithmic process that can be relevant for your explanation. These can be roughly divided in process-based explanations (data selection, model management) and outcome-based explanations.

This guidance helps you to extract the logic/rationale of your decision/model and its outcomes:

1. Identify the type of the algorithm to be explained (Section 1.2 / 1.3)
2. Assess whether the algorithm is directly explainable
 - If required: find supplementary explanation strategy (Section 1.4 / 1.5)
3. Determine non-technical elements of explanation (See general guidance)



**STEP
7**

Implement the technical and organizational measures needed for the explanation

Implement the required technical and organizational measures. Choose methods for interpretability and explainability based on the impact of the algorithmic process and its output, the required type and level of transparency /interpretability, and the relevant audiences.



**STEP
8**

Assess and evaluate

Ensure frequent evaluation of the model explanations by monitoring technical and organization changes relevant to the model outcomes. End-user/subject comprehension of the explanation could also be evaluated to increase the quality of the explanations.

¹ Usually when we talk about explainability, we seem to take a default assumption that we are explaining the machine-made process that led to a given output to the end users. But there are other audiences who want some explanations too. And different audiences would require different levels or types of explanations.

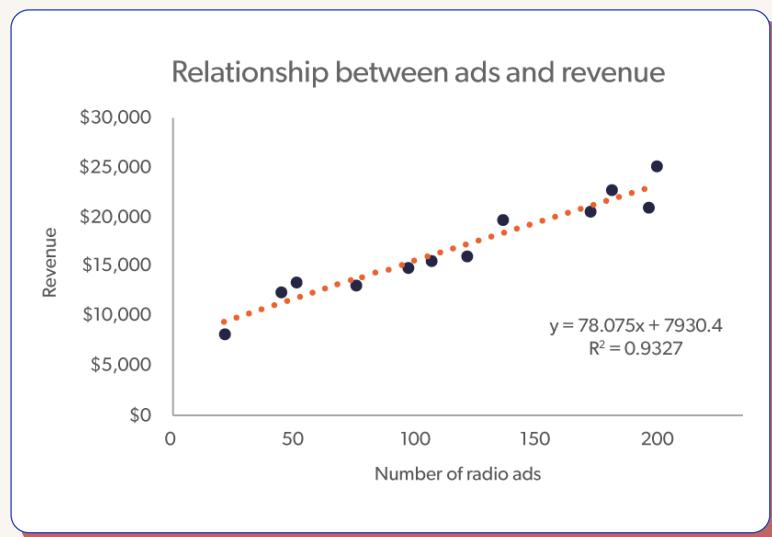
Overview of direct explainable algorithms

This section lists a selection of common algorithms that are directly explainable (with the general disclaimer that an increase in features/dimensions can render a model unexplainable in practice).

1.1.1. Linear regression (LR)

Makes predictions about a (continuous) target variable by summing weighted input/predictor variables. In other words, regression takes one or more independent variables and estimates the relationship with the variable you want to predict. This estimation, often visualized by a regression- or trendline, minimizes the differences between predicted and observed values (in a way that the sum of the deviations is zero). Linear regressions require the dependent and independent variables to be numerical (interval or ratio).²

The regression output, the model, describes the relationship between the dependent and independent variables. For nonlinear relationships, polynomial regressions are used to create a better fit.



Linear regression in Excel showing relationship between Ads and revenue.
Source: <https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/>

Possible Uses

Linear regressions are used in many applications. Its relative simplicity is advantageous in highly regulated sectors like finance (e.g. credit scoring) and healthcare (predict disease risk given e.g. lifestyle and existing health conditions) because it's simpler to calculate and have oversight over.

An example of the use of linear regression in a business setting is to evaluate the relationship between advertising expenditure and revenue.

Interpretability

High level of interpretability because of linearity (variables change at the same rate) and monotonicity (variables change at the same rate). This can become less interpretable with increased number of features (ie high dimensionality) or interactions between features. There are several tools to explore and visualize the regression model and the feature importance beyond the model parameters, see the resources below for guidance.³

² Nominal variables can included using dummy variables ³ For interpreting the model output see: <https://christophm.github.io/interpretable-ml-book/limo.html>

Resources

Type	Name
Primer	Introduction to Linear Regression and Polynomial Regression https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb
Primer	5.1 Linear Regression https://christophm.github.io/interpretable-ml-book/limo.html
Article	Gaines, B. R., & Zhou, H. (2016). Algorithms for fitting the constrained lasso. Journal of Computational and Graphical Statistics, 27(4), 861-871. https://hua-zhou.github.io/media/pdf/GainesKimZhou08CClasso.pdf
Article	Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. http://beehive.cs.princeton.edu/course/read/tibshirani-jrssb-1996.pdf
Tutorial (P)	A beginner's guide to Linear Regression in Python with Scikit-Learn https://medium.com/analytics-vidhya/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-6b0fe70b32d7
Tutorial (P)	Simple and Multiple Linear Regression in Python https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9
Tutorial (R)	Linear Regression http://r-statistics.co/Linear-Regression.html
Tutorial (R)	Linear Regression in R https://www.datacamp.com/community/tutorials/linear-regression-R
Tutorial (P/R)	Interpret R Linear/Multiple Regression output https://medium.com/analytics-vidhya/interpret-r-linear-multiple-regression-output-lm-output-point-by-point-also-with-python-8e53b2ee2a40
Visualization (R)	{jtools} Tools for summarizing and visualizing regression models https://cran.r-project.org/web/packages/jtools/vignettes/summ.html

1.1.2. Logistic regression

Extends linear regression to classification problems by using a logistic function to transform outputs to a probability between 0 and 1. Linear or polynomial regression needs a numerical dependent variable, logistic regression outputs a probability for a class (e.g. this email is spam, this is a picture of a cat). Logistic regression is often considered part of the Generalized Familiar Model (GLM) family (see Section 1.1.40.)

Possible Uses

Like linear regression, advantageous in highly regulated and safety-critical sectors, but in use cases that are based in classification problems such as yes/no decisions on risks, credit, or disease.

An example of the use of logistic regression is spam detection. Using an existing dataset containing emails and the classification if the email is spam, a logistic regression model is trained on a section of the dataset. Once trained (and validated by the data that was not used for training) the model will provide a probability that a message is spam.

Interpretability

Good level of interpretability but less so than linear regression because features are transformed through a logistic function and related to the probabilistic result logarithmically rather than as sums. Interpreting these log-odds can be non-intuitive (see below for a primer). Communicating probabilities to the general public can be a challenge in itself (see resources for guidance).

Resources

Type	Name
Primer	Introduction to Logistic Regression https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
Primer	Understanding Logistic Regression Coefficients https://towardsdatascience.com/understanding-logistic-regression-coefficients-7a719ebcd35
Primer	5.2 Logistic regression https://christophm.github.io/interpretable-ml-book/logistic.html
Tutorial (P)	Building A logistic regression in python https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
Tutorial (R)	Logistic Regression http://r-statistics.co/Logistic-Regression-With-R.html
Primer	Understanding uncertainty: Visualizing probabilities https://plus.maths.org/content/understanding-uncertainty-visualising-probabilities

1.1.3. Regularized regression (LASSO and Ridge)⁴

Extends linear regression (Section 1.1.1) by adding penalization and regularization to feature weights to increase sparsity/reduce dimensionality (i.e. making the model simpler). Regularization aims to resolve overfitting and thus increases the generalizability by discouraging large weights (or even eliminating them from the model). In other words, the resulting model is less biased to the sample it was trained on and less elaborate (more sparse). This makes Lasso models more suitable for relatively small datasets.

Possible Uses

Like linear regression, regularized regression is advantageous in highly regulated and safety-critical sectors that require understandable, accessible, and transparent results. Regularized regression is useful for high dimensional data, for instance in economic/financial forecasting.

Interpretability

High level of interpretability due to improvements in the sparsity of the model through better feature selection procedures. In other words, regularized regression can “automatically” suppress/discount

⁴ Lasso is commonly used for feature selection since it sets some of the feature weights to 0; while ridge regression shrinks overall weights but it doesn't do feature selection.

those correlated features. For correlated features, it will assign higher coefficients to a few of them, and suppress the others (nonetheless, regularized regression probably cannot give very good explainability about how it makes the selection among a group of features; and how the dummy variables are encoded would also impact the regularized regression).

Resources

Type	Name
Primer	Ridge and Lasso Regression: L1 and L2 Regularization https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b
Primer	How to Perform Lasso and Ridge Regression in Python https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression

1.1.4. Generalized linear model (GLM)

GLM "generalizes" linear regressions by allowing the linear model to be related to the dependent variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. This enables linear modelling between variables that do not follow a normal distribution, have discrete values, etc.

Possible Uses

This extension of LR is applicable to use cases where target variables have constraints that require the exponential family set of distributions (for instance, if a target variable involves number of people, units of time or probabilities of outcome, the result has to have a non-negative value).

Interpretability

Good level of interpretability that tracks the advantages of LR while also introducing more flexibility. Because of the link function, determining feature importance may be less straightforward than with the additive character of simple LR, a degree of transparency may be lost.

Resources

Type	Name
Primer	Generalized linear models https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab
Primer	Generalized linear model https://en.wikipedia.org/wiki/Generalized_linear_model
Primer	GLM, GAM and more https://christophm.github.io/interpretable-ml-book/extend-lm.html
Tutorial (R)	Generalized Linear Models https://data.princeton.edu/r/glms
Tutorial (P)	{Statsmodels} Generalized Linear Models https://www.statsmodels.org/stable/glm.html

1.1.5. Generalized additive model (GAM)

Generalized additive models generalize GLMs to include non-linear independent variables. To model non-linear relationships between features and target variables (not captured by LR), a GAM sums non-parametric functions of predictor variables (like splines or tree-based fitting) rather than simple weighted features. GAM can account for non linear and unexpected effects, for instance seasonality in weather models.⁵

Possible Uses

This extension of LR is applicable to use cases where the relationship between predictor and response variables is not linear (i.e where the input-output relationship changes at different rates at different times) but optimal interpretability is desired. An example of GAM use is in times series, for instance modelling financial markets or the weather.⁶

Interpretability

Good level of interpretability because, even in the presence of non-linear relationships, the GAM allows for clear graphical representation of the effects of predictor variables on response variables.

Resources

Type	Name
Primer	GLM, GAM and more https://christophm.github.io/interpretable-ml-book/extend-lm.html
Tutorial (P)	PyGAM https://pygam.readthedocs.io/en/latest/
Tutorial (R)	Generalized Additive Models https://www.r-bloggers.com/generalized-additive-models-ing-probabilities

1.1.6. Decision tree (DT)

DTs are useful with non linear data and where features interact with each other. DTs split a dataset in various subsets according to a cutoff value. DT's moves from starting "root" nodes to terminal "leaf" nodes, following a logical decision path that is determined by Boolean-like "if- then" operators that are weighted through training. To predict the outcome in each "leaf" node, the average of that node is used. How a tree is structured (criteria, number of splits, max levels, etc.) depend on the algorithm used (see the sources below for more).

⁵ Explainable Boosting Machine (EBM) are one of the glass box interpretable algos based on GAMs. <https://github.com/interpretml/interpret> ⁶ See <https://www.kdnuggets.com/2017/04/time-series-analysis-generalized-additive-models.html>



The figure left shows how a descision tree algorithm (CART in this case) splits a dataset and the corresponding decision tree. Source: <https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb>

Possible Uses

Because the step-by-step logic that produces DT outcomes is easily understandable to non-technical users (depending on number of nodes/ features), this method may be used in high-stakes and safety-critical decision- support situations that require transparency as well as many other use cases where volume of relevant features is reasonably low.

An example of an application of Decision Trees is the modelling of eCommerce customer behavior to predict whether a new visitor of a webshop will transact or not based on customer and device attributes, and behavior (time spent on the website, products viewed etc.).⁷

Interpretability

High level of interpretability if the DT is kept manageable small, so that the logic can be followed end-to- end. The weight of the features is clear from the DT. The advantage of DT's over LR is that the former can accommodate non-linearity and variable interaction while remaining interpretable.

Resources

Type	Name
Primer	The Complete Guide to Decision Trees https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14
	Decision tree types https://en.wikipedia.org/wiki/Decision_tree_learning
Article	Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC Press.
Article	Perner, P. (2011, August). How to interpret decision trees?. In Industrial Conference on Data Mining (pp. 40-55). Springer, Berlin, Heidelberg.
Tutorial (R)	A Guide to Machine Learning in R for Beginners: Decision Trees https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb

⁷ See <https://github.com/aamirpatel23/ECommerce-Decision-Trees> for an implementation

Visualization	{Scikit-learn} How to visualize decision trees https://yusout.com/2019/05/07/how-to-visualize-decision-trees/
Tutorial (R)	Linear Regression in R https://www.datacamp.com/community/tutorials/linear-regression-R

1.1.7. Rule/decision lists and sets

Closely related to DT's, rule/decision lists and sets apply a series of if-then statements to input features in order to generate predictions. Whereas decision lists are ordered and narrow down the logic behind an output by applying "else" rules, decision sets keep individual if-then statements unordered and largely independent, while weighting them so that rule voting can occur in generating predictions.

Possible Uses

As with DT's, because the logic that produces rule lists and sets is easily understandable to non-technical users, this method may be used in high-stakes and safety-critical decision-support situations that require transparency as well as many other use cases where the clear and fully transparent justification of outcomes is a priority.

Applications of rule lists can be found in clinical settings, for instance to determine stroke risks:

```

if hemiplegia and age > 60 then stroke risk 58.9% (53.8%—63.8%)
else if cerebrovascular disorder then stroke risk 47.8% (44.8%—50.7%)
else if transient ischaemic attack then stroke risk 23.8% (19.5%—28.4%)
else if occlusion and stenosis of carotid artery without infarction then
stroke risk 15.8% (12.2%—19.6%)
else if altered state of consciousness and age > 60 then stroke risk
16.0% (12.2%—20.2%)
else if age ≤ 70 then stroke risk 4.6% (3.9%—5.4%)
else stroke risk 8.7% (7.9%—9.6%)

```

Source: Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). *Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model*. *The Annals of Applied Statistics*, 9(3), 1350-1371.

Interpretability

Rule lists and sets have one of the highest degrees of interpretability of all optimally performing and non-opaque algorithmic techniques. However, they also share with DT's the same possibility that degrees of understandability are lost as the rule lists get longer or the rule sets get larger.

Resources

Type	Name
Primer	Decision Rules https://christophm.github.io/interpretable-ml-book/rules.html
Package (R)	OneR https://cran.r-project.org/web/packages/OneR/

1.1.8. Supersparse linear integer model (SLIM)

SLIM utilizes data-driven learning to generate a simple scoring system that only requires users to add, subtract, and multiply a few numbers in order to make a prediction. Because SLIM produces such a sparse and accessible model, it can be implemented quickly and efficiently by non-technical users, who need no special training to deploy the system.

Possible Uses

SLIM has been used in medical applications that require quick and streamlined but optimally accurate clinical decision-making. A version called Risk-Calibrated SLIM (RiskSLIM) has been applied to the criminal justice sector to show that its sparse linear methods are as effective for recidivism prediction as some opaque models that are in use.

1. Prior Arrests ≥ 2	1 point	...				
2. Prior Arrests ≥ 5	1 point	+				
3. Prior Arrests for Local Ordinance	1 point	+				
4. Age at Release between 18 to 24	1 point	+				
5. Age at Release ≥ 40	-1 point	+				
SCORE		= ...				
SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

RiskSlim scoring system for recidivism prediction. Source: Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. Interfaces, 48(5), 449-466.

Interpretability

Because of its sparse and easily understandable character, SLIM offers optimal interpretability for human-centered decision-support. As a manually completed scoring system, it also ensures the active engagement of the interpreter-user, who implements it.

Resources

Type	Name
Article	Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. Available at SSRN 2919024. https://arxiv.org/pdf/1702.04690.pdf
Article	Rudin, C., & Ustun, B. (2018). Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. <i>Interfaces</i> , 48(5), 449- 466. https://users.cs.duke.edu/~cynthia/docs/WagnerPrizeCurrent.pdf
Article	Ustun, B., Traca, S., & Rudin, C. (2013). Supersparse linear integer models for interpretable classification. <i>arXiv preprint arXiv:1306.6677</i> .
Python	Optimized scoring systems for classification problems https://github.com/ustunb/slim-python

1.1.9. Naive Bayes

Uses Bayes rule to estimate the probability that a feature belongs to a given class, assuming that features are independent of each other. To classify a feature, the Naive Bayes classifier computes the posterior probability for the class membership of that feature by multiplying the prior probability of the class with the class conditional probability of the feature.

Possible Uses

While this technique is called Naive for the unrealistic assumption of the independence of features, it is known to be very effective. Its quick calculation time and scalability make it good for applications with high dimensional feature spaces. Common applications include spam filtering, recommender systems, and sentiment analysis.

Naive Bayes classifiers are used for text classification, for instance in chatbots to determine whether a user inputs a statement or question; or to classify a review as positive or negative.⁸

Interpretability

Naive Bayes classifiers are highly interpretable, because the class membership probability of each feature is computed independently. The assumption that the conditional probabilities of the independent variables are statistically independent (this is Naive), however, is also a weakness, because feature interactions are not considered.

Resources

Type	Name
Primer	Naive Bayes Classifier https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
Primer	5.7.1 Naive Bayes Classifier https://christophm.github.io/interpretable-ml-book/other-interpretable.html#naive-bayes-classifier
Article	Možina, M., Demšar, J., Kattan, M., & Zupan, B. (2004, September). Nomo-grams for visualization of naive Bayesian classifier. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 337-348). Springer, Berlin, Heidelberg.
Tutorial (P)	{Scikit-learn} 1.9. Naive Bayes https://scikit-learn.org/stable/modules/naive_bayes.html
Tutorial (R)	Understanding Naive Bayes Classifier Using R https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/

1.1.10. K-nearest neighbor (KNN)

Used to group data into clusters for purposes of either classification or prediction, this technique identifies a neighborhood of nearest neighbors around a data point of concern and either finds the mean outcome of them for prediction or the most common class among them for classification.

⁸ See for an example: <https://towardsdatascience.com/unfolding-naive-bayes-from-scratch-2e86dcae4b01>

Possible Uses

KNN is a simple, intuitive, versatile technique that has wide applications but works best with smaller datasets. Because it is non-parametric (makes no assumptions about the underlying data distribution), it is effective for non-linear data without losing interpretability. Common applications include recommender systems, image recognition, anomaly detection⁹, and customer rating and sorting.

Interpretability

KNN works off the assumption that classes or outcomes can be predicted by looking at the proximity of the data points upon which they depend to data points that yielded similar classes and outcomes. This intuition about the importance of nearness/proximity is the explanation of all KNN results. Such an explanation is more convincing when the feature space remains small, so that similarity between instances remains accessible.

Resources

Type	Name
Primer	Machine Learning Basics with the K-Nearest Neighbors Algorithm https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
Tutorial (P)	K Nearest NeighborAlgorithm In Python https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55
Tutorial (R)	K-nearest Neighbors Algorithm with Examples in R (Simply Explained knn) https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c

1.2. Overview of indirectly interpretable algorithms

The algorithms in this section are not directly interpretable. To extract the logic for an outcome, a supplementary explanation has to be generated using the strategies outlined in Section 1.5.

1.2.1. Support vector machines (SVM)

Uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space. In other words, an SVM finds the optimal way to divide a dataset, by focussing on points close to the boundary between two classes. An SVM therefore finds a divider that splits two classes by maximizing the margin (or street) between closest points.

Possible Uses

SVM's are extremely versatile for complex sorting tasks. They can be used to detect the presence of objects in images (face/no face; cat/no cat), to classify text types (sports article/arts article), and to identify genes of interest in bioinformatics.

⁹ For instance to detect credit card fraud, see: Malini, N., & Pushpa, M. (2017, February). Analysis on credit card fraud identification techniques based on KNN and outlier detection. In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) (pp. 255-258). IEEE.

Interpretability

Low level of interpretability that depends on the dimensionality of the feature space. In context-determined cases, the use of SVM's should be supplemented by secondary explanation tools.

SVM can be considered a linear classifier that has the maximum margin from the two classes. The level of interpretability could be actually similar to that of a linear model. Its interpretability decreases when kernel trick is applied to address the non-linear data.

Resources

Type	Name
Primer	A Practical Guide to Interpreting and Visualizing Support Vector Machines https://towardsdatascience.com/a-practical-guide-to-interpreting-and-visualising-support-vector-machines-97d2a5b0564e
Article	Explaining Support Vector Machines: A Color Based Nomogram https://pubmed.ncbi.nlm.nih.gov/27723811/
Tutorial (P)	{Scikit-learn} 1.4. Support Vector Machines https://scikit-learn.org/stable/modules/naive_bayes.html
Tutorial (R)	{libsvm} Support Vector Machines https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf
Visualization	Van Belle, et al. (2016). Explaining Support Vector Machines: A Color Based Nomogram. PloS one, 11(10), e0164568. https://doi.org/10.1371/journal.pone.0164568

1.2.2. Artificial neural net (ANN)

Neural nets are a family of non-linear statistical techniques (including recurrent, convolutional, and deep neural nets) that build complex mapping functions to predict or classify data by employing the feedforward - and sometimes feedback - of input variables through trained networks of interconnected and multi-layered operations.

Possible Uses

ANN's are best suited to complete a wide range of classification and prediction tasks for high dimensional feature ibid. cases where there are very large input vectors. Their uses may range from computer vision, image recognition, sales and weather forecasting, pharmaceutical discovery, and stock prediction to machine translation, disease diagnosis, and fraud detection.

Interpretability

The tendencies towards curviness (extreme non-linearity) and high-dimensionality of input variables produce very low-levels of interpretability in ANN's. They are considered to be the epitome of "black box" techniques. Where appropriate, the use of ANN's should be supplemented by secondary explanation tools (see Supplementary explanation strategies and tools). Consequently, where global interpretation of the model would be required, (most) ANNs will not be compliant.¹⁰

¹⁰ But there are efforts to generate global explanations for neural networks. For instance: Ibrahim, M. et al. (2019, January). Global explanations of neural networks: Mapping the landscape of predictions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp.279-287).

ANN's are often used where hard "evidence" for a decision is unavailable (content moderation for instance), precisely because the task is hard to define (and ANNs are able to extract a signal from seemingly noisy/trivial data).

Resources

Type	Name
Primer	Chapter 10 Neural Network Interpretation https://christophm.github.io/interpretable-ml-book/neural-networks.html
Resources	https://www.kdnuggets.com/2018/12/finlayson-machine-learning-resources.html

1.2.3. Random Forest

Builds a predictive model by combining and aggregating the results from multiple (sometimes thousands) of decision trees that are trained on random subsets of shared features and training data.¹¹

Possible Uses

Random forests are often used to effectively boost the performance of individual decision trees, to improve their error rates, and to mitigate overfitting. They are very popular in high-dimensional problem areas like genomic medicine and have also been used extensively in computational linguistics, econometrics, and predictive risk modelling.

Interpretability

Very low levels of interpretability may result from the method of training these ensembles of decision trees on bagged data and randomized features, the number of trees in a given forest, and the possibility that individual trees may have hundreds or even thousands of nodes.

Resources

Type	Name
Primer	Understanding Random Forest https://towardsdatascience.com/understanding-random-forest-58381e0602d2
Primer	Interpretability and Random Forests https://towardsdatascience.com/interpretability-and-random-forests-4fe13a79ae34
Article	Bénard, C. et al. (2019). SIRUS: Making Random Forests Interpretable. https://arxiv.org/pdf/1908.06852.pdf

¹¹ Random forest provides impurity based feature importance metrics. The higher, the more important the feature. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_

1.2.4. Ensemble methods

As their name suggests, ensemble methods are a diverse class of meta-techniques that combines different "learner" models (of the same or different type) into one bigger model (predictive or classification) in order to decrease the statistical bias, lessen the variance, or improve the performance of any one of the sub-models taken separately. (Gradient) Boosting is one of the widely used ensemble methods.

Possible Uses

Ensemble methods have a wide range of applications that tracks the potential uses of their constituent learner models (these may include DT's, KNN's, Random Forests, Naive Bayes, etc.).

Interpretability

The interpretability of Ensemble Methods varies depending upon what kinds of methods are used. For instance, the rationale of a model that uses bagging techniques, which average together multiple estimates from learners trained on random subsets of data, may be difficult to explain. Explanation needs of these kinds of techniques should be thought through on a case-by-case basis.

Resources

Type	Name
Primer	The Method of Boosting https://www.r-bloggers.com/the-method-of-boosting/
Tutorial (P)	{Scikit-learn} 1.11. Ensemble methods https://scikit-learn.org/stable/modules/ensemble.html
Tutorial (R)	Gradient Boosting and Parameter Tuning in R https://www.kaggle.com/camnugent/gradient-boosting-and-parameter-tuning-in-r
Package (R)	xgboostExplainer https://github.com/AppliedDataSciencePartners/xgboostExplainer

1.3. Supplementary explanation strategies and tools

This section lists some of the strategies for extracting information from your model for explaining a decision. Many of the strategies in this section are part of the toolkits listed in the next section.

The supplementary strategies are model agnostic, they do not depend on the model to extract information. This makes them very flexible as the strategies are applicable to more than one type of model. The strategies all try to uncover the feature importance (and interactions), generate a simpler model, or provide context through counterfactuals.

1.3.1. Surrogate models (SM) [Post-hoc¹², local & global¹³]

SM's build a simpler interpretable model (often a decision tree or rule list) from the dataset and predictions of an opaque system. The purpose of the SM is to provide an understandable proxy of the complex model that estimates that model well, while not having the same degree of opacity. They are good for assisting in processes of model diagnosis and improvement and can help to expose overfitting and bias. They can also represent some non-linearities and interactions that exist in the original model.

Limitations

As approximations, SM's often fail to capture the full extent of non-linear relationships and high-dimensional interactions among features. There is a seemingly unavoidable trade-off between the need for the SM to be sufficiently simple so that it is understandable by humans, and the need for that model to be sufficiently complex so that it can represent the intricacies of how the mapping function of a "black box" model works as a whole. That said, the R2 (the proportion of variance explained by the model) measurement can provide a good quantitative metric of the accuracy of the SM's approximation of the original complex model.

Resources

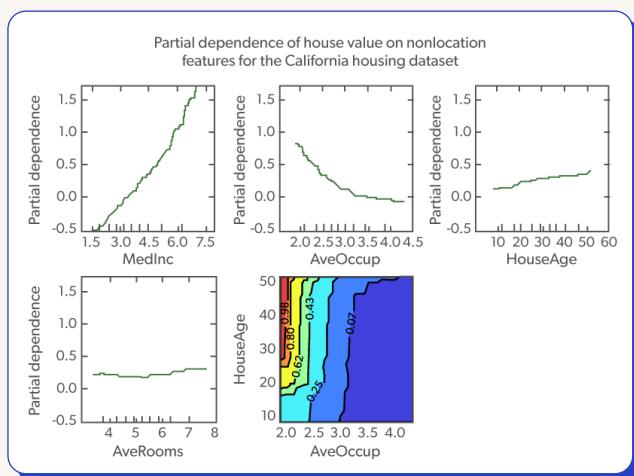
Type	Name
Primer	8.6 Global Surrogate https://christophm.github.io/interpretable-ml-book/global.html
Article	Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction. https://arxiv.org/abs/1706.09773
Article	Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In Advances in neural information processing systems (pp. 24-30). http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf
Article	Van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In European Conference on Machine Learning (pp. 418-429). Springer, Berlin, Heidelberg. https://link.springer.com/content/pdf/10.1007/978-3-540-74958-5_39.pdf
Article	Valdes, G., Luna, J. M., Eaton, E., Simone II, C. B., Ungar, L. H., & Solberg, T. D. (2016). MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. <i>Scientific reports</i> , 6, 37854. https://www.nature.com/articles/srep37854

¹² Post hoc interpretability refers to the application of interpretation methods after model training, while intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models. This criteria distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after training (post hoc), in <https://christophm.github.io/interpretable-ml-book/taxonomy-of-interpretability-methods.html>. ¹³ "Interpretability" can be broadly divided into global interpretability, meaning understanding the entirety of a trained model including all decision paths, and local interpretability, the goal of understanding the results of a trained model on a specific input and small deviations from that input. In "Assessing the Local Interpretability of Machine Learning Models", Dylan Slack, Sorelle A. Friedler, Carlos Scheidegger, Chiradeep Dutta Roy.

Python toolbox	Surrogate Modeling Toolbox https://github.com/SMTorg/smt
Package (R)	{SPOT} https://cran.r-project.org/web/packages/SPOT/SPOT.pdf

1.3.2. Partial Dependence Plot (PDP) [Post-hoc global]

A PDP calculates and graphically represents the marginal effect of one or two input features on the output of an opaque model by probing the dependency relation between the input variable(s) of interest and the predicted outcome across the dataset, while averaging out the effect of all the other features in the model. This is a good visualization tool, which allows a clear and intuitive representation of the nonlinear behavior for complex functions (like random forests and SVM's). It is helpful, for instance, in showing that a given model of interest meets monotonicity constraints across the distribution it fits.



Partial dependence plot in scikit-learn.
Source: https://scikit-learn.org/0.18/auto_examples/ensemble/plot_partial_dependence.html

Limitations

While PDP's allow for valuable access to non-linear relationships between predictor and response variables, and therefore also for comparisons of model behavior with domain-informed expectations of reasonable relationships between features and outcomes, they do not account for interactions between the input variables under consideration. They may, in this way, be misleading when certain features of interest are strongly correlated with other model features.

Because PDP's average out marginal effects, they may also be misleading if features have uneven effects on the response function across different subsets of the data - i.e. where they have different associations with the output at different points. This means that PDP's can provide a general understanding of feature importance, but cannot serve as an individual explanation.

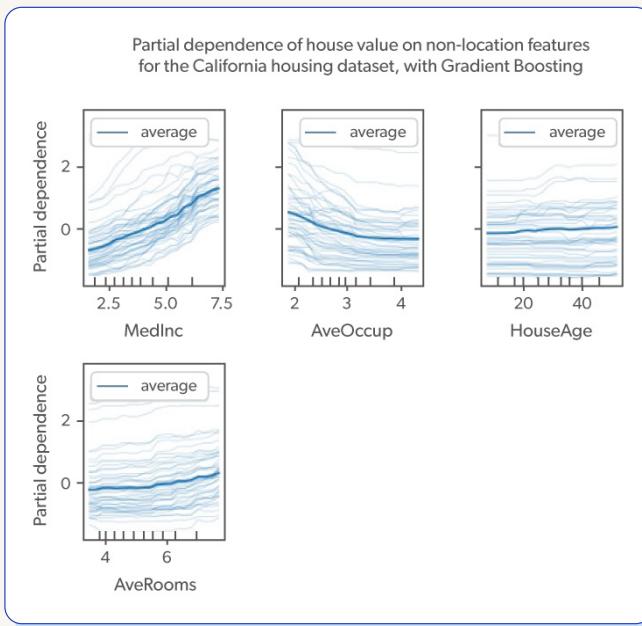
Type	Name
Article	Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. <i>Annals of statistics</i> , 1189-1232. https://projecteuclid.org/download/pdf_1/euclid-aos/1013203451

Resources

Article	Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. <i>The R Journal</i> , 9(1), 421-436. https://journal.r-project.org/archive/2017/RJ-2017-016/RJ-2017-016.pdf
Package (R)	pdp: Partial Dependence Plots https://cran.r-project.org/web/packages/pdp/index.html
Primer	8.1 Partial Dependence Plot (PDP) https://christophm.github.io/interpretable-ml-book/pdp.html
Python	PDPbox https://github.com/SauceCat/PDPbox
Primer	Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) https://christophm.github.io/interpretable-ml-book/ice.html

1.3.3. Individual Conditional Expectations Plot (ICE) [Post-hoc local]

Refining and extending PDP's, ICE plots graphs of how the predictions change when a feature changes per instance. Significantly, ICE plots therefore disaggregate or break down the averaging of partial feature effects generated in a PDP by showing changes in the feature-output relationship for each specific instance, i.e. observation-by- observation. This means that it can both detect interactions and account for uneven associations of predictor and response variables.



Partial Dependence (thick line) and Individual Conditional expectations (thin lines) in scikit-learn
Source: https://scikit-learn.org/dev/auto_examples/inspection/plot_partial_dependence.html#id6

Limitations

When used in combination with PDP's, ICE plots can provide local information about feature behavior that enhances the coarser global explanations offered by PDP's. Most importantly, ICE plots are able to detect the interaction effects and heterogeneity in features that remain hidden from PDP's¹⁴ in virtue of the way they compute the partial dependence of outputs on features of interest by averaging out the effect of the

¹⁴ Because PDP's average out.

other predictor variables. Still, although ICE plots can identify interactions, they are also liable to missing significant correlations between features and become misleading in some instances.

Constructing ICE plots can also become challenging when datasets are very large. In these cases, time-saving approximation techniques such as sampling observation or binning variables can be employed (but, depending on adjustments and size of the dataset, with an unavoidable impact on explanation accuracy).

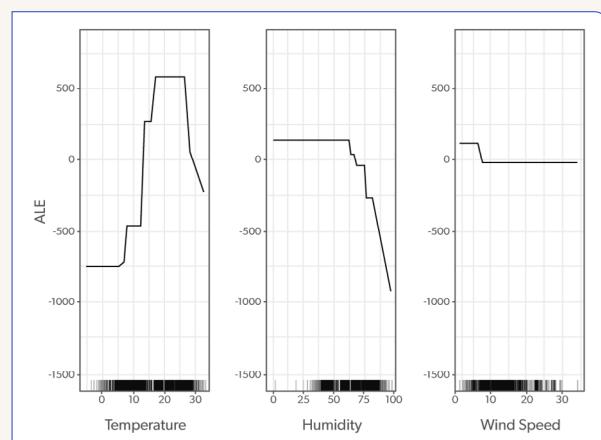
Resources

Type	Name
Article	Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. <i>Journal of Computational and Graphical Statistics</i> , 24(1), 44-65. https://arxiv.org/pdf/1309.6392.pdf
Package (R)	Gam: Generalized Additive Models https://CRAN.R-project.org/package=gam
Book	https://christophm.github.io/interpretable-ml-book/ice.html
Python	https://github.com/AustinRochford/PyCEbox

1.3.4. Accumulated Local Effects Plots (ALE) [Post-hoc global]

As an alternative approach to PDP's, ALE plots provide a visualization of the influence of individual features on the predictions of a "black box" model by averaging the sum of prediction differences for instances of features of interest in localized intervals and then integrating these averaged effects across all of the intervals. By doing this, they are able to graph the accumulated local effects of the features on the response function as a whole. Because ALE plots use local differences in prediction when computing the averaged influence of the feature (instead of its marginal effect as do PDP's), it is able to better account for feature interactions and avoid statistical bias. This ability to estimate and represent feature influence in a correlation-aware manner is an advantage of ALE plots.

ALE plots are also more computationally tractable than PDP's because they are able to use techniques to compute effects in smaller intervals and chunks of observations.



ALE plot showing the effects of three features in a model predicting bike rentals.
Source: <https://christophm.github.io/interpretable-ml-book/ale.html>

Limitations

A notable limitation of ALE plots has to do with the way that they carve up the data distribution into intervals that are largely chosen by the explanation designer. If there are too many intervals, the prediction differences may become too small and less stably estimate influences. If the intervals are widened too much, the graph will cease to sufficiently represent the complexity of the underlying model. While ALE plots are good for providing global explanations that account for feature correlations, the strengths of using PDP's in combination with ICE plots should also be considered (especially when there are less interaction effects in the model being explained). All three visualization techniques shed light on different dimensions of interest in explaining opaque systems, so the appropriateness of employing them should be weighed case-by-case.

Resources

Type	Name
Article	Apley, D. W., & Zhu, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468. https://arxiv.org/pdf/1612.08468.pdf
Package (R)	https://cran.r-project.org/web/packages/ALEPlot/index.html
Book	https://christophm.github.io/interpretable-ml-book/ale.html
Python	https://github.com/blent-ai/ALEPython

1.3.5. Global Variable Importance [Post-hoc global]

The global variable importance strategy calculates the contribution of each input feature to model output across the dataset by permuting the feature of interest and measuring changes in the prediction error; if changing the value of the permuted feature increases the model error, then that feature is considered to be important. Utilising global variable importance to understand the relative influence of features on the performance of the model can provide significant insight into the logic underlying the model's behavior. This method also provides valuable understanding about non-linearities in the complex model that is being explained.¹⁵

Limitations

While permuting variables to measure their relative importance, to some extent, accounts for interaction effects, there is still a high degree of imprecision in the method with regard to which variables are interacting and how much these interactions are impacting the performance of the model.

A bigger picture limitation of global variable importance comes from what is known as the "Rashomon effect". This refers to the variety of different models that may fit the same data distribution equally well. These models may have very different sets of significant features. Because the permutation-based technique can only provide explanatory insight with regard to a single model's performance, it is unable to address this wider problem of the variety of effective explanation schemes.

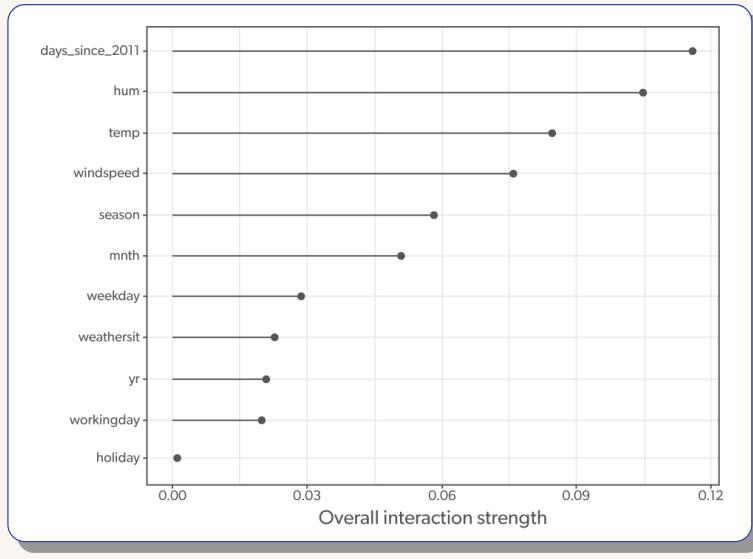
¹⁵ Global variable importance is also called permutation feature importance

Resources

Type	Name
Article	Breiman, L. (2001). Random forests. Machine learning, 45(1), p.5-32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf
Article	Casalicchio, G., Molnar, C., & Bischl, B. (2018, September). Visualizing the feature importance for black box models. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 655-670). Springer, Cham. https://arxiv.org/pdf/1804.06620.pdf
Article	Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv:1801.01489 https://arxiv.org/abs/1801.01489
Article	Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. arXiv preprint arXiv:1801.01489. https://arxiv.org/abs/1801.01489v2
Article	Hooker, G., & Mentch, L. (2019). Please Stop Permuting Features: An Explanation and Alternatives. arXiv preprint arXiv:1905.03151. https://arxiv.org/pdf/1905.03151.pdf
Article	Zhou, Z., & Hooker, G. (2019). Unbiased Measurement of Feature Importance in Tree-Based Methods. arXiv preprint arXiv:1903.05179. https://arxiv.org/pdf/1903.05179.pdf
Package (R)	Random Uniform Forests for Classification, Regression and Unsupervised Learning https://cran.r-project.org/web/packages/randomUniformForest/index.html

1.3.6. Global Variable Interaction [Post-hoc global]

The global variable interaction strategy computes the importance of variable interactions across the dataset by measuring the variance in the model's prediction when potentially interacting variables are assumed to be independent. This is primarily done by calculating an "H-statistic" where a no-interaction partial dependence function is subtracted from an observed partial dependence function in order to compute the variance in the prediction. This is a versatile explanation strategy, which has been employed to calculate interaction effects in many types of complex models including ANN's and Random Forests. It can be used to calculate interactions between two or more variables and also between variables and the response function as a whole. It has been effectively used, for example, in biological research to identify interaction effects among genes.



*Global Interaction of features in a bike rental model
 Source: <https://christophm.github.io/interpretable-ml-book/interaction.html>*

Limitations

While the basic capacity to identify interaction effects in complex models is a positive contribution of global variable interaction as a supplementary explanatory strategy, there are a couple of potential drawbacks to which you may want to pay attention.

First, there is no established metric in this method to determine the quantitative threshold across which measured interactions become significant. The relative significance of interactions is useful information as such, but there is no way to know at which point interactions are strong enough to exercise effects.

Second, the computational burden of this explanation strategy is very high, because interaction effects are being calculated combinatorially across all the data points. This means that as the number of data points increases, the number of necessary computations increases exponentially.

Resources

Type	Name
Article	Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. <i>The Annals of Applied Statistics</i> , 2(3), 916-954. https://projecteuclid.org/download/pdfview_1/euclid.aoas/1223908046
Article	Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. <i>arXiv preprint arXiv:1805.04755</i> . https://arxiv.org/pdf/1805.04755.pdf
Tutorial (P)	Hooker, G. (2004, August). Discovering additive structure in black box functions. In <i>Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining</i> (pp. 575-580). ACM. https://dl.acm.org/doi/10.1145/1014052.1014122
Primer	https://christophm.github.io/interpretable-ml-book/interaction.html

1.3.7. Sensitivity Analysis and Layer-Wise Relevance Propagation (LRP) [Post-hoc local (possibly global)]

Sensitivity analysis and LRP are supplementary explanation tools used for artificial neural networks. Sensitivity analysis identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output's sensitivity to such changes in input values identifies the most relevant features. LRP is another method to identify feature relevance that is downstream from sensitivity analysis. It uses a strategy of moving backward through the layers of a neural net graph to map patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.



Heatmap showing relevant pixels for the classification of a digit.

Notice how the "top" and the "end of the curl" are relevant for classification of an image as the digit 5.

Source: <https://towardsdatascience.com/in-depth-layer-wise-relevance-propagation-340f95deb1ea>

Limitations

Both sensitivity analysis and LRP identify important variables in the vastly large feature spaces of neural nets. These explanatory techniques find visually informative patterns by mathematically piecing together the values of individual nodes in the network. As a consequence of this piecemeal approach, they offer very little by way of an account of the reasoning or logic behind the results of an ANNs' data processing.

Recently, more and more research has focused on attention-based methods of identifying the higher-order representations that are guiding the mapping functions of these kinds of models as well as on interpretable CBR methods that are integrated into ANN architectures and that analyse images by identifying prototypical parts and combining them into a representational wholes. These newer techniques are showing that some significant progress is being made in uncovering the underlying logic of some ANN's.

Resources

Type	Name
Toolkit	Skater https://oracle.github.io/Skater/overview.html
Toolkit	DeepLIFT https://github.com/kundajelab/deeplift

1.3.8. Local Interpretable Model-Agnostic Explanation (LIME) and anchors [Post-hoc local]

LIME works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account

for the features which figure prominently in the specific prediction or classification under scrutiny. LIME does this by generating a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is locally faithful to that instance. Note that other interpretive models like decision trees may be used as well.

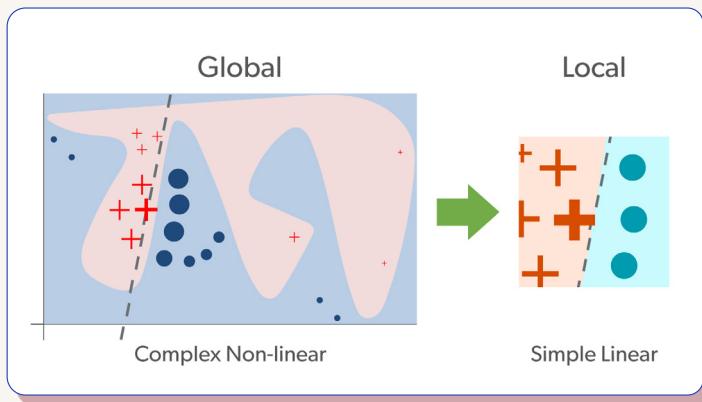


Diagram showing LIME strategy of generating a simple local linear model from a global complex and non-linear model.
Source: <https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html>

Limitations

While LIME appears to be a step in the right direction, in its versatility and in the availability of many iterations in very usable software, a host of issues that present challenges to the approach remain unresolved.

For instance, the crucial aspect of how to properly define the proximity measure for the "neighborhood" or "local region" where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable, even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified interpretable model that successfully approximates the underlying model reasonably well near any given data point.

LIME's creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call "anchors". These "high precision rules" incorporate into their formal structures "reasonable patterns" that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.

Resources

Type	Name
Article	Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM. https://arxiv.org/pdf/1602.04938.pdf
Python	https://github.com/marcotcr/lime
Package (R)	https://cran.r-project.org/web/packages/lime/index.html

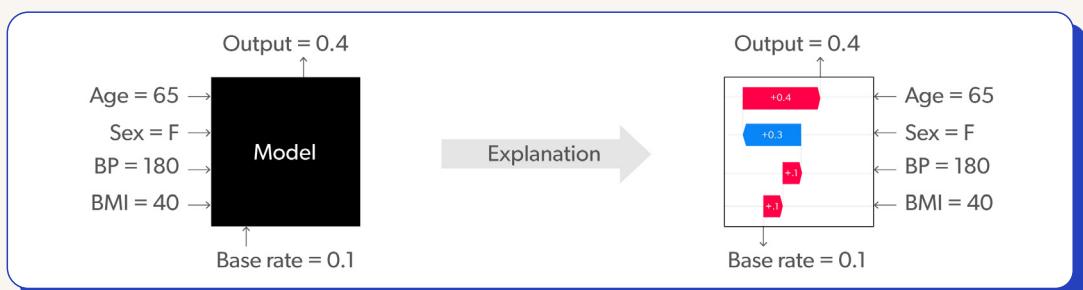
Article	Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/11491
Python	https://github.com/marcotcr/anchor
Tutorial (R)	Explaining the Explainer: A First Theoretical Analysis of LIME; https://arxiv.org/abs/2001.03447

1.3.9. Shapley Additive ExPlanations (SHAP) [local post hoc]

SHAP uses concepts from cooperative game theory to define a "Shapley value" for a feature of concern that provides a measurement of its influence on the underlying model's prediction.

Broadly, this value is calculated by averaging the feature's marginal contribution to every possible prediction for the instance under consideration. The way SHAP computes marginal contributions is by constructing two instances: the first instance includes the feature being measured, while the second leaves it out by substituting a randomly selected stand-in variable for it. After calculating the prediction for each of these instances by plugging their values into the original model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be computed.

This method then allows SHAP, by extension, to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. While computationally intensive, this means that for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. This computational robustness has made SHAP attractive as an explainer for a wide variety of complex models, because it can provide a more comprehensive picture of relative feature influence for a given instance than any other post-hoc explanation tool.



Shapley values (right) to explain a black box model.

Source: <https://github.com/slundberg/shap>

Limitations

Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold.

Note, though, some later SHAP versions do offer methods of approximation such as Kernel SHAP and Shapley Sampling Values to avoid this excessive computational expense. These methods do, however, affect the overall accuracy of the method.

Another significant limitation of SHAP is that its method of sampling values in order to measure marginal variable contributions assumes feature independence (i.e. that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between the stand-in variables that are used as substitutes for left-out features are necessarily unaccounted for when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced, because the complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating.

There are currently efforts being made to account for feature dependencies in the SHAP calculations. The original creators of the technique have introduced Tree SHAP to, at least partially, include feature interactions. Others have recently introduced extensions of Kernel SHAP.

Resources

Type	Name
Article	Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
Package (R)	https://modeloriented.github.io/shapper/ https://cran.r-project.org/web/packages/iml/index.html
Python	https://github.com/slundberg/shap

1.3.10. Counterfactual Explanation [Post-hoc local]

Counterfactual explanations offer information about how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realized by the recipient of a particular decision or outcome.

Incorporating counterfactual explanations into a model at its point of delivery allows stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. For AI systems that assist decisions about changeable human actions (like loan decisions or credit scoring), incorporating counterfactual explanation into the development and testing phases of model development may allow the incorporation of actionable variables, ie input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome.

In this way, counterfactual explanatory strategies can be used as a way to incorporate reasonableness and the encouragement of agency into the design and implementation of AI systems.

Limitations

While counterfactual explanation offers a useful way to contrastively explore how feature importance may influence an outcome, it has limitations that originate in the variety of possible features that may be included when considering alternative outcomes. In certain cases, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of possible explanations seem potentially arbitrary.

Moreover, there are as yet limitations on the types of datasets and functions to which these kinds of explanations are applicable.

Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and questionable covariate relationships that may be buried deep within the model's architecture. It is a good idea to use counterfactual explanations in concert with other supplementary explanation strategies - that is, as one component of a more comprehensive explanation portfolio.

Resources

Type	Name
Article	Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066-4076). http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf
Article	Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency(pp. 10-19). ACM. https://arxiv.org/pdf/1809.06514.pdf
Article	Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.
Python	Python https://github.com/ustunb/actionable-recourse
Package (R)	ContrastiveExplanation (Foil Trees) https://github.com/MarcelRobeir/ContrastiveExplanation
Package (R)	DiCE https://github.com/microsoft/DiCE

1.3.11. Self-Explaining and Attention-Based Systems

Self-explaining and attention-based systems actually integrate secondary explanation tools into the opaque systems so that they can offer runtime explanations of their own behaviors. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an "attention-directing" mechanism translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user.

Research into integrating "attention-based" interfaces is continuing to advance toward potentially making their implementations more sensitive to user needs, explanation-forward, and humanly understandable. Moreover, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them.

Limitations

Automating explanations through self-explaining systems is a promising approach for applications where users benefit from gaining real-time insights about the rationale of the complex systems they are operating. However, regardless of their practical utility, these kinds of secondary tools will only work as well as the explanatory infrastructure that is actually unpacking their underlying logics. This explanatory layer must remain accessible to human evaluators and be understandable to affected individuals. Self-explaining systems, in other words, should themselves remain optimally interpretable. The task of formulating a primary strategy of supplementary explanation is still part of the process of building out a system with self-explaining capacity.

Another potential pitfall to consider for self-explaining systems is their ability to mislead or to provide false reassurance to users, especially when humanlike qualities are incorporated into their delivery method. This can be avoided by not designing anthropomorphic qualities into their user interface and by making uncertainty and error metrics explicit in the explanation as it is delivered.

Resources

Type	Name
Article	Tutorial on Attention-based Models (Part 1) https://krntneja.github.io/posts/2018/attention-based-models-1
Tutorial (R)	Attention-based Neural Machine Translation with Keras https://blogs.rstudio.com/ai/posts/2018-07-30-attention-layer/

Explanation toolkits and frameworks

The following section lists a selection of explanation toolkits and frameworks in three sections:

Tools for extraction and explaining models (model)

Tools for detecting bias (model/data)

Tools for designing/visualizing (UI/UX)

Tools for extracting model explanations

Overview

Type	Name	Techniques	Model
DrWhy.ai	R	<ul style="list-style-type: none"> – Model adapters – Model agnostic explainers – Model specific explainers – Automated exploration 	Model agnostic
Alibi	Python (The Integrated Gradients implementation supports only TF/Keras models)	<ul style="list-style-type: none"> – Accumulated Local Effects – Anchors – Contrastive Explanation Method – Counterfactual Instances – Counterfactuals Guided by Prototypes – Guided Integrated Gradients – Kernel SHAP and Tree SHAP – Measuring the linearity of machine learning models – Trust Scores 	Mostly model agnostic. Integrated Gradients is for neural networks only.
Skater	Python	<ul style="list-style-type: none"> – Model agnostic Feature Importance – Model agnostic Partial Dependence Plots – Local Interpretable Model Explanation(LIME) – Layer-wise Relevance Propagation (e-LRP): image – Integrated Gradient: image and text – Scalable Bayesian Rule Lists – Tree Surrogates 	Model agnostic
tf-explain	Python (Tensorflow)	<ul style="list-style-type: none"> – Activations Visualization – Vanilla Gradients – Gradients*Inputs – Occlusion Sensitivity – Grad CAM – SmoothGrad – Integrated Gradients – group fairness metrics derived from selection rates and error rates including rich subgroup fairness – Comprehensive set of sample distortion metrics – Generalized Entropy Index (Differential Fairness and Bias Amplification) 	Neural networks

iml	R	<ul style="list-style-type: none"> – Feature importance – Partial dependence plots – Individual conditional expectation plots – Accumulated local effects – Tree surrogate – LocalModel: Local Interpretable Model-agnostic Explanations – Shapley value for explaining single predictions 	Model agnostic
iNNvestigate	Python (keras)	<ul style="list-style-type: none"> – Function – Signal – Attribution 	Neural Network
treeinterpreter	Python (scikit-learn)	<ul style="list-style-type: none"> – Bias and feature contribution 	Decision tree Random forest
Captum	Python (pytorch)	<ul style="list-style-type: none"> – Attribution using various algorithms, e.g.: – Integrated Gradients – SHAP – GradCAM – LRP – Feature Permutation – Feature Ablation – NoiseTunnel – Layer Attribution – Neuron attribution 	Neural network in PyTorch
DeepExplain	Python (Tensorflow)	<ul style="list-style-type: none"> – Saliency maps – Integrated Gradients – DeepLIFT – E-LRP – Perturbation-based methods – Shapley Value sampling 	Neural networks
grad-cam	Python & R	<ul style="list-style-type: none"> – Saliency mapping 	Neural networks
keras-vis	Python (keras)	<ul style="list-style-type: none"> – Activation maximization – Saliency maps – Class activation maps 	Neural networks
Interpret-ml		<ul style="list-style-type: none"> – Explainable Boosting – Decision Tree – Decision Rule List – Linear/Logistic Regression – SHAP Kernel Explainer – SHAP Tree Explainer – LIME – Morris Sensitivity Analysis – Partial Dependence 	Model agnostic Linear model Decision tree

1.3.12. DrWhy.AI (including DALEX)

The DrWhy.AI universe is a collection of tools for visual Exploration, explanation and debugging of predictive models in R. Packages in the DrWhy.AI family of models may be divided into four classes:

Model adapters

Predictive models created with different tools have different structures, and different interfaces. Model adapters create uniform wrappers. This way other packages may operate on models in an unified way. DALEX is a lightweight package with a generic interface. DALEXtra is a package with extensions for heavyweight interfaces like scikitlearn, h2o, mlr.

Model agnostic explainers

These packages implement specific methods for model exploration. They can be applied to a single model or they can compare different models. ingredients implements variable specific techniques like Ceteris Paribus, Partial Dependency, Permutation based Feature Importance. iBreakDown implements techniques for variable attribution, like Break Down or SHAPley values. auditor implements techniques for model validation, residual diagnostic and performance diagnostic.

Model specific explainers

These packages implement model specific techniques. randomForestExplainer implements techniques for exploration of randomForest models. EI implements techniques for exploration of gbm and xgboost models. cr19 implements techniques for exploration of survival models.

Automated exploration

These packages combine a series of model exploration techniques and produce an automated report of website for model exploration. modelStudio implements a dashboard generator for local and global interactive model exploration. modelDown implements a HTML website generator for global model cross comparison.

Packages fo DrWhy.AI include:

DALEX

(Descriptive mAchine Learning EXplanations) helps to understand how complex models are working. The main function explain creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers.

DALEXtra

is an extension pack for DALEX. This package provides easy to use connectors for models created with scikitlearn, keras, H2O, mljar and mlr.

ingredients

is a collection of tools for assessment of feature importance and feature effects.

iBreakDown

is a model agnostic tool for explanation of predictions from black boxes ML models. Break Down Table shows contributions of every variable to a final prediction. Break Down Plot presents variable contributions in a concise graphical way. SHAP (Shapley Additive Attributions) values are calculated as average from random Break Down profiles. This package works for binary classifiers as well as regression models.

auditor

is a tool for model-agnostic validation. Implemented techniques facilitate assessing and comparing the goodness of fit and performance of models. In addition, they may be used for the analysis of the similarity of residuals and for the identification of outliers and influential observations.

vivo

helps to calculate instance level variable importance (measure of local sensitivity). The importance measure is based on Ceteris Paribus profiles and can be calculated in eight variants.

randomForest-Explainer

helps to understand what is happening inside a Random Forest model. This package helps to explore main effects and pairwise interactions, depth distribution, conditional responses and feature importance.

rSafe

is a model agnostic tool for making an interpretable white-box model more accurate using alternative black-box model called surrogate model. Based on the complicated model, such as neural network or random forest, new features are being extracted and then used in the process of fitting a simpler interpretable model, improving its overall performance.

xspliner

is a collection of tools for training interpretable surrogate ML models. The package helps to build simple, interpretable models that inherits informations provided by more complicated ones - resulting model may be treated as explanation of provided black box, that was supplied prior to the algorithm.

shapper

is an R wrapper of SHAP python library. It accesses python implementation through reticulate connector.

drifter

is an R package that identifies concept drift in model structure or in data structure.

Type	Link
Github	https://github.com/ModelOriented/DrWhy
Book	https://pbiecek.github.io/ema/
Library (P)	https://pypi.org/project/dalex/

1.3.13. Alibi

Alibi is an open source Python library aimed at machine learning model inspection and interpretation. The initial focus on the library is on black-box, instance-based model explanations.

Currently the following methods are supported:

- Anchors
- Contrastive Explanation Method
- Counterfactual Instances
- Counterfactuals Guided by Prototypes
- Guided Integrated Gradients (<https://github.com/samzabdiel/XAI>)
- Kernel SHAP
- Measuring the linearity of machine learning models
- Trust Scores

Type	Link
Github	https://github.com/SeldonIO/alibi
Documentation	https://docs.seldon.io/projects/alibi/en/latest/

1.3.14. Skater

Skater is an open source unified framework for python to enable Model Interpretation for all forms of model to help one build an Interpretable machine learning system often needed for real world use-cases. Skater supports algorithms to demystify the learned structures of a black box model both globally(inference on the basis of a complete data set) and locally(inference about an individual prediction).

Included algorithms are:

- Model agnostic Feature Importance
- Model agnostic Partial Dependence Plots
- Local Interpretable Model Explanation(LIME)
- Layer-wise Relevance Propagation (e-LRP): image
- Integrated Gradient: image and text
- Scalable Bayesian Rule Lists
- Tree Surrogates

Type	Link
Documentation	https://oracle.github.io/Skater/overview.html
API Documentation	https://oracle.github.io/Skater/api.html

1.3.15. tf-explain

tf-explain offers interpretability methods for Tensorflow 2.0 to ease neural network's understanding. With either its core API or its tf.keras callbacks, you can get feedback on the training of your models.

Available methods are:

- Activations Visualization Visualize how a given input comes out of a specific activation layer
- Vanilla Gradients Visualize gradients on the inputs towards the decision.
- Gradients*Inputs Variant of Vanilla Gradients ponderating gradients with input values.
- Occlusion Sensitivity Visualize how parts of the image affects neural network's confidence by occluding parts iteratively
- Grad CAM Visualize how parts of the image affects neural network's output by looking into the activation maps
- SmoothGrad Visualize stabilized gradients on the inputs towards the decision.
- Integrated Gradients Visualize an average of the gradients along the construction of the input towards the decision.

Type	Link
API Documentation	tf-explain https://tf-explain.readthedocs.io/en/latest/
Tutorial (P)	https://gilberttanner.com/blog/interpreting-tensorflow-model-with-tf-explain

1.3.16. iml

iml is an R package that interprets the behavior and explains predictions of machine learning models. It implements model-agnostic interpretability methods - meaning they can be used with any machine learning model.

Methods included are:

- Feature importance
- Partial dependence plots
- Individual conditional expectation plots (ICE)
- Accumulated local effects
- Tree surrogate
- LocalModel: Local Interpretable Model-agnostic Explanations
- Shapley value for explaining single predictions

Type	Link
Documentation	https://github.com/christophM/iml

1.3.17. iNNvestigate

The iNNvestigate library contains implementations for the following methods:

- function:
 - gradient: The gradient of the output neuron with respect to the input.
 - smoothgrad: SmoothGrad averages the gradient over a number of inputs with added noise.
- signal:
 - deconvnet: DeConvNet applies a ReLU in the gradient computation instead of the gradient of a ReLU.
 - guided: Guided BackProp applies a ReLU in the gradient computation additionally to the gradient of a ReLU.
 - pattern.net: PatternNet estimates the input signal of the output neuron.
- attribution:
 - input_t_gradient: Input * Gradient
 - deep_taylor[.bounded]: DeepTaylor computes for each neuron a rootpoint, that is close to the input, but which's output value is 0, and uses this difference to estimate the attribution of each neuron recursively.
 - pattern.attribution: PatternAttribution applies Deep Taylor by searching root points along the singal direction of each neuron.
 - lrp.*: LRP attributes recursively to each neuron's input relevance proportional to its contribution of the neuron output.
 - integrated_gradients: IntegratedGradients integrates the gradient along a path from the input to a reference.
 - deeplift.wrapper: DeepLIFT (wrapper around original code, slower) computes a backpropagation based on "finite" gradients.

Type	Link
Documentation	https://innvestigate.readthedocs.io/en/latest/index.html
Article	Alber, M. et al.(2019). iNNvestigate neural networks. Journal of Machine Learning Research, 20(93), 1-8.

1.3.18. treeinterpreter

Package for interpreting scikit-learn's decision tree and random forest predictions. Allows decomposing each prediction into bias and feature contribution components.

Type	Link
Documentation	https://github.com/andosa/treeinterpreter
Article	http://blog.datadive.net/interpreting-random-forests/

1.3.19. Captum

Captum is a model interpretability and understanding library for PyTorch. Captum contains general purpose implementations of integrated gradients, saliency maps, smoothgrad, vargrad and others for PyTorch models. It has quick integration for models built with domain-specific libraries such as torchvision, torchtext, and others. Captum has been expanded to support adversarial robustness, concept-based interpretability such as TCAV and a number of metrics that measure how trustworthy feature importance scores are.

Type	Link
Documentation	https://github.com/pytorch/captum
Tutorials	https://captum.ai/tutorials/

1.3.20. CausalML

Causal ML is a Python package that provides a suite of uplift modeling and causal inference methods using machine learning algorithms. It provides a standard interface that allows user to estimate the Conditional Average Treatment Effect (CATE) or Individual Treatment Effect (ITE) from experimental or observational data.

Type	Link
Documentation	https://github.com/uber/causalml
Documentation	https://causalml.readthedocs.io/en/latest/about.html

1.3.21. DeepExplain

DeepExplain provides a unified framework for state-of-the-art gradient and perturbation-based attribution methods. It can be used by researchers and practitioners for better understanding the recommended existing models, as well for benchmarking other attribution methods. It supports Tensorflow as well as Keras with Tensorflow backend. Support for PyTorch is planned.

Implements the following methods:

- Gradient-based attribution methods
 - Saliency maps
 - Gradient * Input
 - Integrated Gradients
 - DeepLIFT, in its first variant with Rescale rule (*)
 - -LRP (*)
- Perturbation-based attribution methods
 - Occlusion, as an extension of the grey-box method by Zeiler et al.
 - Shapley Value sampling

1.3.22. grad-cam

Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say "dog" in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. This technique is implemented in various other toolkits as well.

Type	Link
Article	Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
Overview of implementations	https://github.com/topics/grad-cam

1.3.23. Keras-vis

keras-vis is a high-level toolkit for visualizing and debugging your trained keras neural net models.

Currently supported visualizations include:

- Activation maximization
- Saliency maps
- Class activation maps

All visualizations by default support N-dimensional image inputs. i.e., it generalizes to N-dim image inputs to your model. The toolkit generalizes all of the above as energy minimization problems with a clean, easy to use, and extendable interface. Compatible with both theano and tensorflow backends with "channels_first", "channels_last" data format.

Type	Link
Documentation	https://github.com/raghakot/keras-vis
Documentation	https://raghakot.github.io/keras-vis/

1.3.24. Interpret-ml

InterpretML is an open-source python package that incorporates state-of-the-art machine learning interpretability techniques under one roof. With this package, you can train interpretable glassbox models and explain blackbox systems. InterpretML helps you understand your model's global behavior, or understand the reasons behind individual predictions.

- Explainable Boosting
- Decision Tree
- Decision Rule List
- Linear/Logistic Regression
- SHAP Kernel Explainer
- SHAP Tree Explainer
- LIME
- Morris Sensitivity Analysis
- Partial Dependence

Type	Link
Documentation	https://github.com/interpretml/interpret
Website	https://interpret.ml/

Tools for detecting Bias

1.3.25. AI Fairness 360

IBM's toolkit to examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education.

Developers note that: the toolkit should only be used in a very limited setting: allocation or risk assessment problems with well-defined protected attributes in which one would like to have some sort of statistical or mathematical notion of sameness. Even then, the code and collateral contained in AIF360 is only a starting point to a broader discussion among multiple stakeholders on overall decision-making workflows.

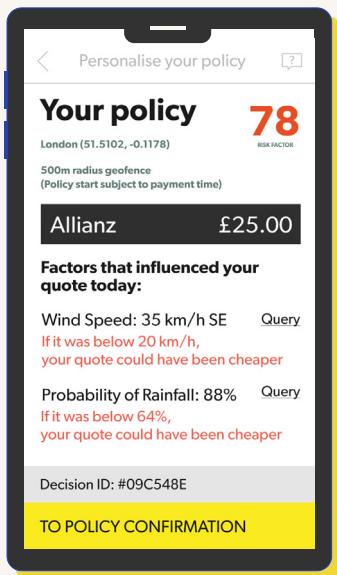
Type	Link
Documentation	https://github.com/IBM/AIF360
API Documentation	https://aif360.readthedocs.io/en/latest/
Article	Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." IBM Journal of Research and Development 63.4/5 (2019): 4-1. https://arxiv.org/pdf/1810.01943.pdf

1.3.26. ML Fairnes-gym

ML-fairness-gym is a set of components for building simple simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments. ML-fairness-gym implements a generalized framework for studying and probing long term fairness effects in carefully constructed simulation scenarios where a learning agent interacts with an environment over time.

Type	Link
Documentation	https://github.com/google/ml-fairness-gym
Blog	https://ai.googleblog.com/2020/02/ml-fairness-gym-tool-for-exploring-long.html

1.4. Examples UX design



Example model explanation for drone insurance
Source: <https://automated-decisions.tumblr.com>

1.4.1. Projects by If

This organization provides toolkits and several blogs related to AI transparency. They also provide a data patterns catalogue, including a section on understanding automated decisions.

Type	Link
Patterns catalog	https://catalogue.projectsbyif.com
Blogs	https://automated-decisions.tumblr.com

1.4.2. TTC Labs

TTC labs provide blogs and a toolkit on designing for trust, transparency, control. Many designs generated in so-called "design jams" are accessible. They provide a toolkit with methodologies and tools for discovery, ideation, prototyping, design ideas and research.

Type	Link
Toolkit	https://toolkit.ttclabs.net/
Designs	https://www.ttclabs.net/designs

Glossary of terms

Class	Category of a feature. For instance the category "Cat" of feature "Pets"
Classification	Identification of which two or more categories a case falls under. For instance: "Item1 is a Cat"
Clusters label	The descriptive, human readable name of clusters identified through a clustering algorithm.
Decision boundary	Divides the decision space in two or more sets. The classifier classifies points in the same set as the same class. Whether the boundary is binary or fuzzy (the transition from each class is discontinuous or gradual) depends on the specific model.
Dimensionality	Refers to the number of features used in the model. Dimensionality (reduction) has a large effect on the quality of the model.
Distribution	A description of the relation of the values and frequency of those values for a feature. Several "archetype" distributions exist. Many statistical techniques assume a variable has a specific underlying distribution (parametric).
Feature	The individual measurable property of a phenomenon, the input/independent variable.
Feature (or problem) space	The "space" formed by the feature (vectors) in the analysis, i.e. all possible values for a set of variables. This spatial representation of data enables measurements of "closeness" and by extension classification.
Generalizability	The models ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
Linearity and non linearity	Linearity means that if one of two related elements changes a little, the other changes a little as well. In a nonlinear relationship, one element does not change in direct proportion to a change in the other element.
Gradient (descent)	Refers to the rate of ascent (or change) of an independent value in relation to changes of a dependent variable. For model optimization "gradient descent" is an optimization strategy to find local minimums in highly dimensional feature spaces.
Logistic function	Function represented by a sigmoid (or s-shaped) curve. Logistic functions are used to map continuous input variables to binary outcome variables.

Glossary of terms

Monotonicity	Refers to the stable relation of two variables, if one variable changes the other always changes in the same way (the rate might vary). I.e. a U-shaped curve does not describe a monotonic relationship, an exponential curve does describe a monotonic relationship.
Scedasticity	The distribution of error terms. Homoscedasticity (the error terms are random and with constant variance) is a common assumption of statistical techniques. Wrongfully assuming homoscedasticity makes model estimates less precise and makes confidence intervals less accurate.
Sparsity	Refers to the limitation of features in the model (lowering dimensions). Adding many features to a model can be problematic for methods that require statistical significance (more dimensions means data become sparse). Additionally, high dimensional models are more complex and harder to explain.