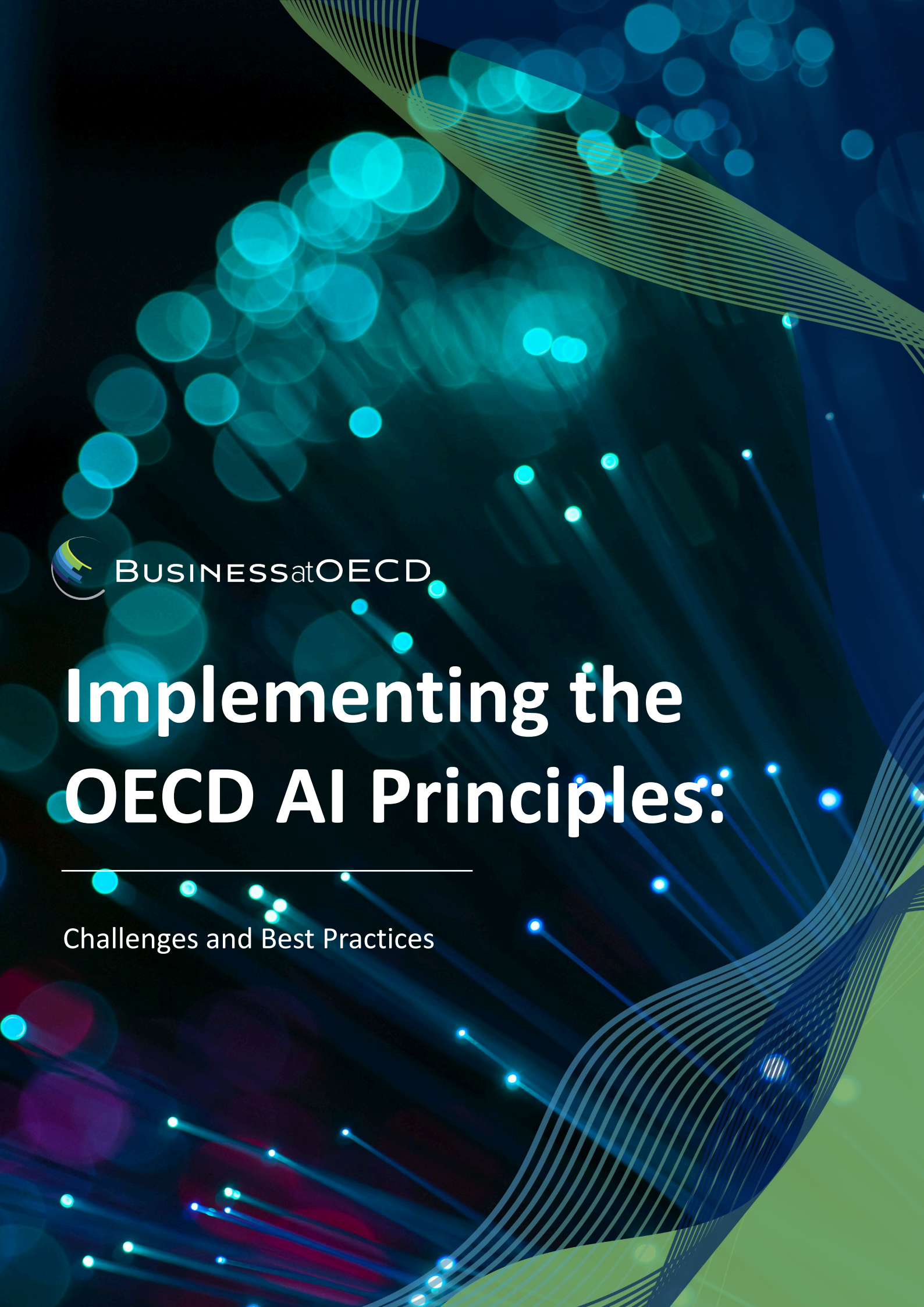


Implementing the OECD AI Principles:

Challenges and Best Practices

Acknowledgements

This project "Implementing the OECD AI Principles: Challenges and Best Practices" is a *Business at OECD* contribution to the work of ONE.TAI to develop a framework and a database of tools as a reference for governments, businesses and civil society to ensure the development of AI systems in line with the OECD Principles for trustworthy AI. This work has been led by *Business at OECD* Committee on Digital Economy Policy under the guidance of Nicole Primmer and Maylis Berviller from the *Business at OECD* Secretariat.

We would like to thank participating *Business at OECD* Members and the OECD for their support of this project, including the development of this Report, and the associated Project Roundtable.

We extend a special thanks to Theodoros Evgeniou, Pal Boza, for serving as consultants on this project, responsible for the development of the seven business case studies and drafting this analytical report.

Business at OECD (BIAC) Report on Implementing the OECD AI Principles: Challenges and Best Practices

July 5, 2022

In 2021, *Business at OECD* (BIAC) launched a project to facilitate the adoption, dissemination and implementation of the OECD AI Principles. By evaluating concrete examples where businesses aimed to implement the Principles, the project complements efforts by the OECD to implement AI in a way that is ethical, lawful, robust and respectful of human rights and democratic values.

The project rests on two pillars: in-depth qualitative research around case studies in order to identify the learnings and challenges companies face when implementing the OECD AI Principles; and evaluation of specific AI systems using the OECD's database of tools for trustworthy AI . The results of our work also contribute to the OECD.AI database of tools for trustworthy AI, good practices and lessons learned from private sector AI initiatives. We are pleased to present the results of this project in the following report, including seven in depth business case studies.

Contents

Introduction:
Business implementation of the OECD AI Principles

On May 22nd 2019, forty-two OECD and partner countries adopted the first intergovernmental policy guidelines on Artificial Intelligence (OECD AI Principles) with the objective to provide guidance for the development of trustworthy AI systems.

Subsequently in 2019, following the launch of its AI Observatory platform, the OECD established the OECD Network of Experts on AI (ONE.AI), convening multistakeholder expert groups addressing trustworthy AI, the classification of AI systems, and national AI policies to deliver evidence and analysis related to critical aspects of AI systems alongside taxonomies for implementation of the OECD AI Principles.

In this context, the OECD Network of Experts Working Group on Implementing Trustworthy AI (ONE.TAI) has been working to move "from principles to practice" through the development of a framework and a database of tools as a reference for governments, businesses, and civil society to ensure the development of AI systems in line with the OECD Principles for trustworthy AI. As stated by the ONE.TAI Chairs, "the framework helps AI practitioners determine which tool fits their use case and how well it supports the OECD AI Principles for trustworthy AI".

With the expected growth of the OECD database of tools in the coming years, the goal of the framework and database is to ensure that organizations make effective use of this important resource and guidance for best practice. In the same way that the successful adoption of new technologies depends on several technical, social and organizational factors, the successful adoption of tools for the development of trustworthy AI requires attention to multi-faceted issues. Sharing and learning from best practices can significantly increase the successful implementation of trustworthy AI – thereby ensuring the overall adoption of safe and responsible AI.

Based on this conclusion, *Business at OECD* (BIAC) launched a project to further study business best practices and related challenges faced in the development of AI tools and processes. The project draws conclusions from a series of business case studies carried out with *Business at OECD* member companies highlighting both the development and use of tools strengthening trustworthy AI.[1] The seven cases were selected to ensure coverage of the five OECD values-based Principles for Trustworthy AI, with each business case describing tools used to implement specific OECD AI Principles. Table 1 provides a summary of the seven cases, tools and principles studied.[2]

To carry out this project, we enlisted the expertise of INSEAD professor Theodoros Evgeniou[3], and researcher Pal Boza[4] to work with seven member companies, who were invited to share and discuss a specific practice - be it a product, tool, process or design methodology - and map it to one of the OECD's AI Principles. This allowed us to assess what impact a specific AI-related practice might have on the operational implementation of the corresponding AI principle.

Companies that participated in the study include *AXA*, *Amazon Web Services* (*AWS*), *IBM*, *Facebook/Meta*, *Microsoft*, *NEC* and *PricewaterhouseCoopers* (*PwC)*. We would like to thank the experts from participating companies who contributed to the development of the case studies.

The tools presented in this study tackle a wide range of AI related challenges, which organizations are facing when developing, deploying, and using AI in a responsible way. They ensure that AI ultimately creates value for business and society, whilst managing the risks that may result from their development and deployment.

Table 1: Case studies and the Trustworthy AI principles

Case-study	Main OECD value-based principles
AWS	Human-centered Values and Fairness; Robustness, Security and Safety
AXA	Fairness
Meta	Transparency and Explainability
IBM	Transparency and Explainability; Accountability
NEC	Robustness, Security and Safety; Accountability
Microsoft	Transparency and Explainability
PwC	Inclusive Growth, Sustainable Development and Well-being

Each of the seven case studies focus on at least one of the OECD AI Principles:

- An important framework from *AWS* provides guidance on how to critically think about all important aspects of responsible AI;
- *AXA*'s tool explores how fairness can be implemented in ways that are best aligned with the context where AI is applied;
- *Meta*'s (formerly *Facebook*) tool addresses the subject of explainability and transparency;
- *IBM*'s contribution discusses how transparency can support AI accountability;
- The discussion with *NEC* shows how AI systems quality assurance and robustness can be achieved, building also on strong and time-tested software product quality management principles;
- *Microsoft*'s Responsible AI tool helps organizations to better understand how a given responsible AI system can be developed and work;
- The toolkit developed by *PwC* includes a set of technical assets and important management as well as governance frameworks to assess and plan for safe and responsible AI development and adoption.

The case studies were based on a series of structured interviews that were designed around five key dimensions. This ensured consistency and allowed the space to develop general lessons across all the cases. These five dimensions addressed the "what", "how", and "who" of the tool, as well as the main challenges faced and best practices that resulted from the process. The study sought to illustrate these aspects of the case by walking through a short example of the specific tool in action. The full text of the seven cases studies appear in the Appendix of the report.

In developing the case studies as a framework, we considered the following sets of questions:

1. What is the AI tool or practice in question?

- *What does the tool aim to achieve?*
- *Why did you decide to develop a tool in this area?*
    - *What were the motivating forces?*
    - *What are the business implications*

2. How does the tool work?

- *At what point in the AI lifecycle should the tool be used?*
- *How is the tool embedded or integrated in the AI product development processes (i.e., how does it assist and inform the user of the tool)?*
- *How does the tool support the continuous monitoring of an AI product while in use?*

3. Who is using the tool during the development of AI products?

- *What are their roles and responsibilities?*
- *Does the tool require specific skills or training to use?*
    - *If so, how is the training conducted?*
- *Does the implementation and use of the tool need the support of specialists, or can it be used by all once developed?*
- *Is the tool business-facing or customer-facing?*
- *Is it usable by other companies or sectors?*

4. What is an example when you used the proposed tool to implement a particular AI principle?

5. What are the challenges, limitations, and best practices to consider while developing or using the tool?

- *What recommendations do you have for regulators as they design the upcoming AI regulations?*

Whilst the case studies provided numerous lessons and examples in their own right, analyzing them as an ensemble through a similar lens highlighted common threads that make for some important learnings. The focus of the following analysis illustrates learnings gained through the practical implementation of AI tools and processes, designed and deployed in line with the OECD Trustworthy AI Principles. In the following section, we share learnings from best practices and challenges related to the tools presented, which subsequently call for increased sharing of knowledge regarding the use of tools to implement trustworthy AI.

Key challenges and best practices in using tools to implement AI Principles

Our project provides several insights and lessons intending to inform a range of actors within the AI ecosystem, including executives, AI developers, deployers, users, and managers, as well as governments and regulators. These insights highlight important factors related to the digital readiness of an organization and users when adopting AI tools or processes; the inclusion of diverse stakeholders; communication practices; organizational buy-in; and change in the development and implementation of AI tools and processes.

Successful AI adoption depends on the deliberate prioritization of good AI governance

Responsible AI policy and tools need to be well aligned with the organization's larger governance structure and its general mission. If not, it may be challenging to successfully implement the tools into practice. Proper AI governance needs to be embedded at all levels of an organization, with clear channels of communication and escalation regarding potential AI risks.

The commitment of an organization's policy team is especially crucial as they share the joint responsibility, together with other teams (e.g., management teams; technical AI teams), to implement trustworthy AI systems. The real change happens, however, when companies allocate full-time personnel to AI governance in the organization. This prevents suboptimal trade-offs regarding how people with other pressing tasks, of which AI governance may just be one, allocate their time in dealing with AI governance issues, and avoid inefficient outcomes.

There is a careful balance to be ensured between standardization of corporate practices and customization of individual solutions based on the AI system's specific context

Across multiple interviews, it was made clear that "there is no one methodology fits all solution" with respect to several different tools. For example, *Meta* (formerly *Facebook*) shared that in the initial development phase of their "Why Am I Seeing This" (WAIST) tool, they expected the exercise to be the same across all case studies. When this did not turn out to be the case, a different methodology was finally developed for each case. At the same time, this case also highlights the importance of a consistent and standardized design. A similar design was used across applications of WAIST on *Facebook* (e.g., advertisement, connected content, recommended content). This approach helped users better understand the tool and made adoption easier.

Balancing standardization and context-dependent customization is key. Appropriate customization of the tool to each project increases precision and effectiveness since AI system requirements can differ for each case study. However, some fundamental requirements appear identical for all AI development projects and may benefit from some standardization. Human oversight and judgement on overwriting the guideline of the tool, in these instances, can provide flexibility and help customization. This flexibility is provided based on a case-by-case decision of the product managers to eventually overwrite the guidelines if needed to adapt to the special circumstances of each project, while informing the stakeholders, including clients, about possible related risks. In this instance, a process to decide and communicate any diversions from standard practice needs to be in place, even in case of diversions from checklists like with the *NEC* tool. Using clear and relevant examples in the form of a flowchart can help in deciding when to divert from the standardized process, as illustrated by the *AXA* Fairness Compass. Similarly, *IBM*'s FactSheets encourages a user-centric methodology for AI transparency that tailors the documentation to the specifics of each case study.[5]

Bear in mind that transparency does not automatically equate to explainability

A key learning from *Meta*'s example on explainability and transparency is that transparency, in itself, does not lead to explainability. Good explanations include signals and factors used by the algorithms most relevant and comprehensible to users. Counter to the common understanding that precision is essential to transparency, explanations may need to leave space for approximations. Systematic use of precise and scientific terms essential to program an algorithm, coupled with a barrage of signals and factors without a rubric to give them sense might not be in the interest of, nor useful to a general audience. In particular when explaining the reasoning behind a specific outcome, approximations about the complex functioning of the algorithm best serve the twin goals of transparency and explainability in a balanced way.

It is also important to recognize that transparency itself can take different forms, and therefore a robust decision-making process should be in place to choose which algorithm(s) to focus on for the given context. Outcome-based transparency tools provide explanations about the specific outputs of an algorithm, while process-based tools focus on how decisions, such as content selection, are made in general, possibly including information about the inputs used. For example, within those types of transparency, *Meta* considers three approaches: (a) tools-based transparency – meaning transparency provided through the user interface; (b) prose-based transparency – transparency provided through written information or explanations, and (c) data-based transparency – information provided to the users about the *Meta* algorithms' content recommendation and ranking choices.

Upskill teams by providing appropriate training in both technical and non-technical aspects of AI

As mentioned previously, working with diverse teams is critical for tackling the complex issues related to AI governance and the development of responsible AI, which sit at the intersection of several disciplines. Due to having multiple people involved in developing an AI tool and ensuring its good governance, specific skills and a certain level of digital and AI literacy are required. This can be ensured by designing and offering tailored training based on educational backgrounds, experiences, and the different roles in the organization. For example, *NEC* made education a priority with a set of courses called "NEC Academy for ALL". In this program, several types of expertise are classified according to each phase of AI usage. For each of these phases, custom training programs have been developed to upskill the relevant teams. Alongside this internal project, *NEC* was also involved in the Japanese government initiative on AI education, "Japan Inter-University Consortium for Mathematics & Data Science Education".

It is also important to note that data literacy is not only about technical training. Other issues pertaining to the societal and ethical aspects of developing and deploying AI are just as important. For example, ensuring that teams are aware of the different types of biases potentially present in a model or dataset is critical, as well as understanding and assessing different types of risks associated with a model. This type of training can be imperative to help teams assess when a model can be considered acceptable from both a technical standpoint, but also and most importantly from a responsibility perspective.

Ensure that there are structures in place to efficiently secure wider organizational buy-in for the development and deployment of an AI tool

Any new tool requires buy-in from the larger organization to be successfully adopted and continually used. Securing this buy-in can be a complex exercise and may require a significant investment of time and effort at all levels of the organization. *AWS*, for example, developed a tool through a structured process of alignment, "accept it, digest it, improve it", which was key for their success as a large company. Similarly, *IBM* noticed that largely involving stakeholders was key in creating their FactSheets. This allowed them to solicit well-balanced feedback and avoid an overly technology-driven process.

Interestingly, certain tools themselves can help with organizational buy-in, internal communications, and coordination. *AWS*'s tool, for example, can help facilitate buy-in on behalf of top-level management by providing a framework for discussion in non-technical terms, which is helpful for the efficient adoption of responsible use principles in organizations.

Finally, building on existing tools and knowledge during the development of an AI tool can be beneficial, especially to build trust and acceptance of the tool by the organization as there is already a precedent in place, for example for engineers who would use the tool.

Ensure continuous improvement

Continuous improvement and update of the tools are key to ensure the relevance of AI technologies because of the fast-paced evolution that is characteristic of AI driven solutions. This is not just the case for the end-product or tool, but also needs to be prioritized for the tools used in the development process of the end AI products and services. For example, a current limitation of the Fairness Compass of *AXA* is that it is only adapted to classification problems, and does not help with the further technical implementation of the fairness metrics. Commissioning ongoing research to adapt the tool to more complex regression problems, amongst others, is a good way to mitigate this issue. Furthermore, *AXA*'s tool will be enhanced through the development of a dedicated library of technical methods related to each potential outcome of the tool's decision tree. This process of continual updating will allow them to achieve the desired statistical fairness metrics (e.g., adversarial approaches, weighing approaches, etc.), which can be later implemented by data scientists.

Soliciting and acting on feedback from both stakeholders outside the organization (e.g., clients using the tool) as well as those inside (e.g., internal users of the tool) is a good way for continuous improvement. In fact, continuous feedback from users and other specific audiences is crucial – both during and after AI systems development. The perspectives of different user groups, academics, and civil society can only be understood through constant consultation and feedback. Having exchanges with experts to fully understand the possible societal implications of tools, as well as the practical benefits of the AI principles one tries to implement (e.g., the meaning of AI explainability tools for polarization, civic society, etc.), is imperative.

Key considerations for regulators

While the focus of the project remains on facilitating the successful development, adoption, and usage of tools to implement trustworthy AI, several lessons can be drawn from the case studies that can serve as useful considerations for regulators.

Allow for different approaches to achieve the implementation of AI principles.

- For example, regulators increasingly consider transparency and explainability as regulatory requirements for AI systems. However, for transparency and explainability to be most effective, regulators need to work with industry experts in the development of any guidance on how to satisfy transparency, increase explainability, and offer control to users.
- Striking a balance between these three goals may be challenging without some flexibility to use different approaches. Providing information about transparency activities can be a good solution. For example, one option can be to let companies demonstrate what they do for transparency through documenting transparency practices and processes.
- Regulatory guidance and positions on transparency also need to be formulated according to their specific audience, since user groups can be very different (e.g., from a digital literacy point of view).

Complement high level requirements with industry best practices and technical tools to ensure AI fairness.

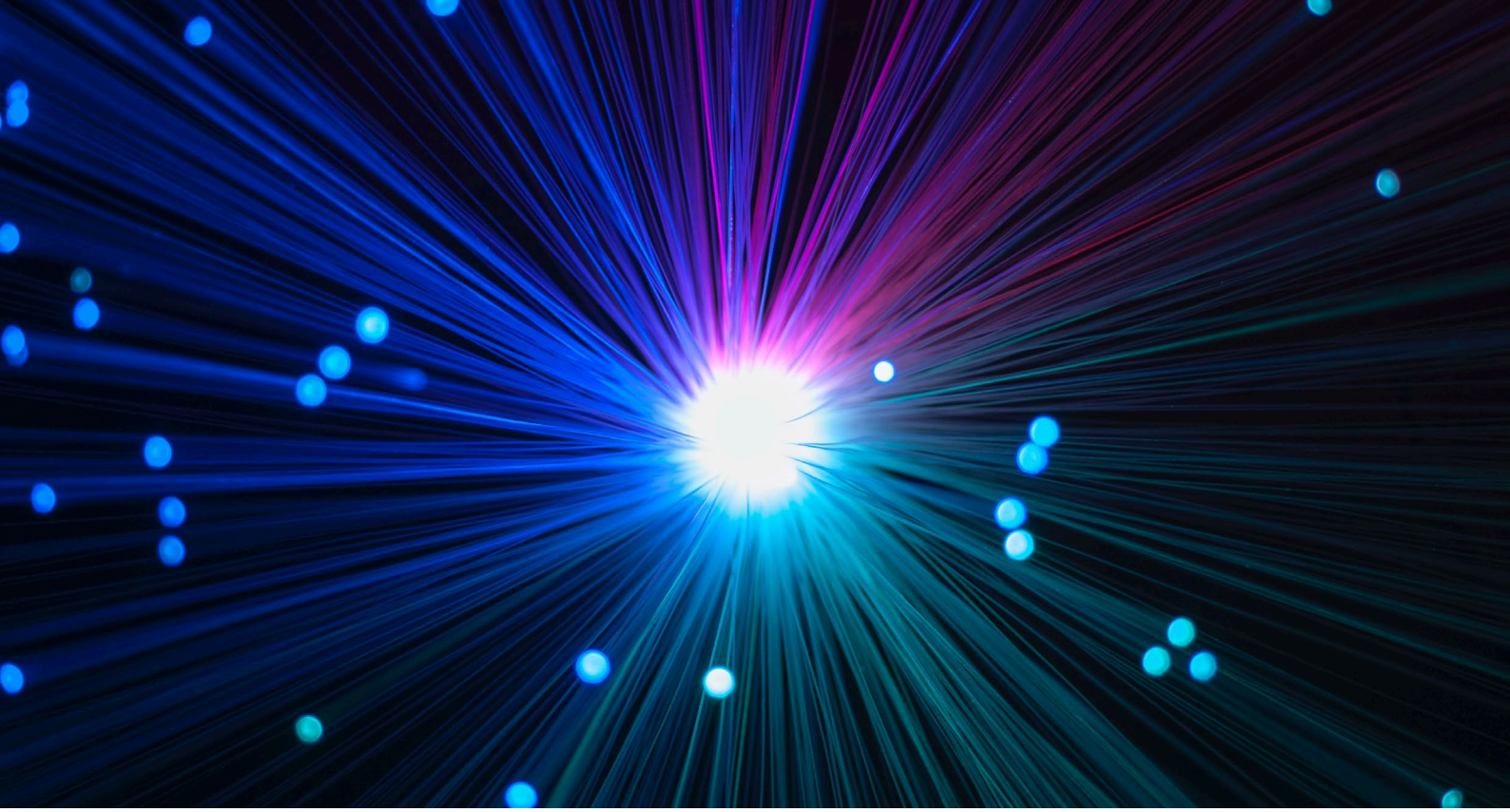
- Fairness is a complex issue, and it can have multiple definitions depending on the context.
- Complementing high-level requirements with industry best practices and technical tools would be an effective way to ensure the development and deployment of AI products and systems possible.

Consider how limitations on the collection and sharing of sensitive data (e.g., gender, age, etc.) can hinder the creation of fair AI solutions.

- Algorithms can learn biases through different proxies, for example, insurance-related data embed possible biases that the driver of certain types of cars may be more likely to be female.
- Regulatory frameworks which allow companies to use sensitive data while restricting their public access is an important consideration in the effort to address and eliminate bias in AI systems. For example, the “AI Act” proposed by European Union Commission, aims to allow the collection and use of personal data for developing AI systems in the public interest. In this case, some third-party entities could have the legal authorization to collect sensitive information to evaluate AI models developed by organizations in each industry.

Look to established AI tools and best practices as important sources of information and complement legislation. This important added value of AI tools themselves provides important information for regulators.

Consider that existing standards (e.g., published by IEEE, ISO, etc.) are becoming increasingly important tools in the regulators’ playbook. Major international and national standardization bodies are working on AI standards to effectively draw on the expertise of industry and formulate best practices to ensure the development and deployment of responsible AI within organizations.



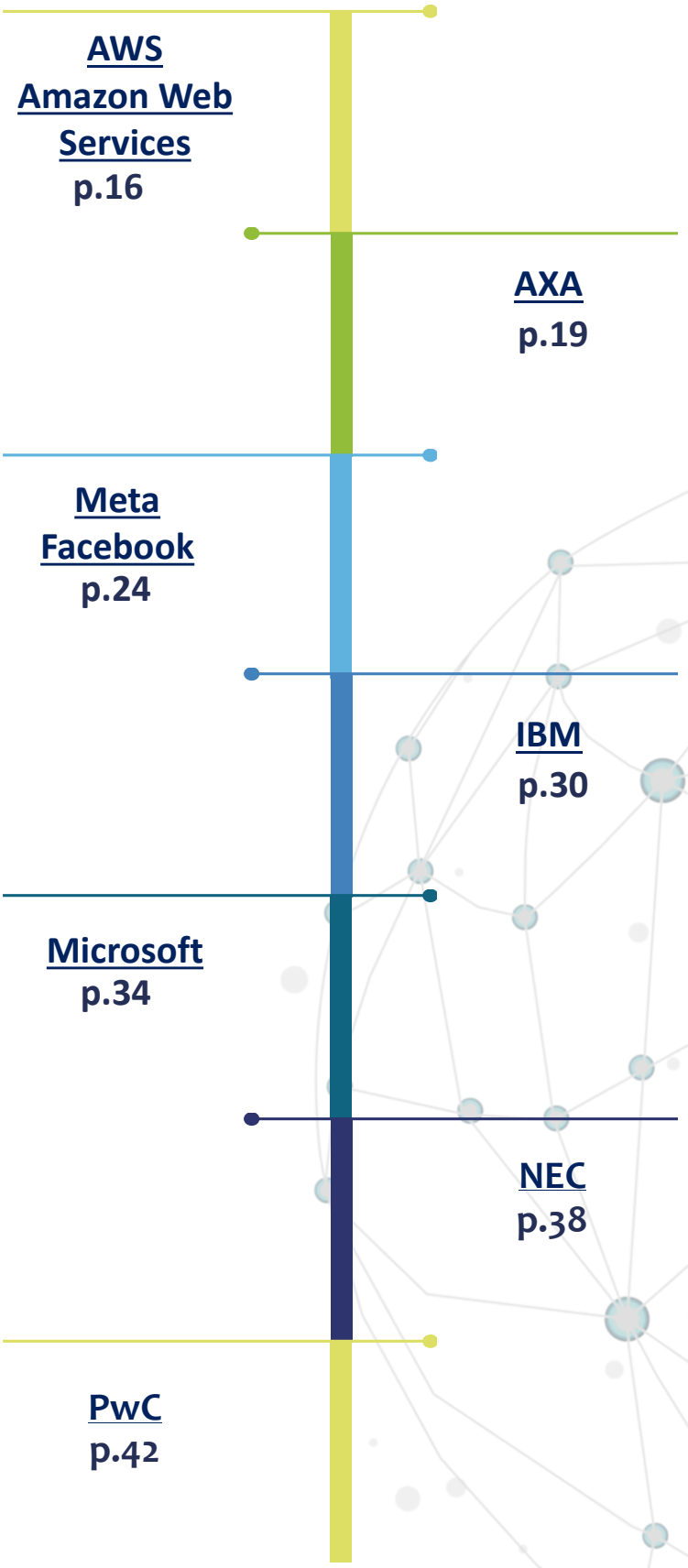
Moving Forward

Implementing trustworthy AI requires not only awareness and availability of tools one can use for this purpose, but also a proper understanding of their operation to best leverage these tools' potential to achieve the best possible outcomes. At the same time, the extent to which regulators grasp the numerous challenges and tradeoffs faced by organizations is critical when drafting future regulations in order to ensure their successful implementation by the private sector. Both objectives have been at the core of this *Business at OECD* project.

This project also highlights that the adoption and successful use of tools to implement AI principles can have significant side effects, impacting the overall culture of organizations, markets and society. Responsible AI tools can have positive impacts in addition to ensuring the quality of AI systems themselves, such as increasing organizational engagement, collaboration, but also helping communication about AI systems across the organization.

The journey to implement AI principles can prove to be much more impactful, but also more complex, than anticipated. We hope that the case studies and lessons learned from this project will help advance the development of trustworthy AI, including to informing the debate around AI regulation, provide useful insight regarding the management of risks related to the development of future AI systems, and support development of best practices regarding the practical implementation of Trustworthy AI across sectors.

Annex
Company case studies



Responsible use of Machine Learning systems tool at Amazon Web Services (AWS)

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

Amazon Web Services (*AWS*) is one of the largest global cloud providers, offering fully featured services supported by around 200 data centres. *AWS* has over 100.000 customers using Machine Learning (ML) on their systems, covering a wide variety of case studies. While the company considers that all use of ML must respect the rule of law, human rights, and values of equity, privacy, and fairness, its customers are also asking for guidance on how to responsibly develop and use ML systems. In this context, *AWS*'s Responsible ML tool has been created based on the following considerations:

- In recent years, several international organisations, governments, academic institutions, as well as businesses have worked on developing principles and tools to guide for responsible ML systems. This has led to the initiation of both "higher-level" recommendations (e.g., OECD AI Principles) and "lower-level" very specific and focused technical tools (e.g., measures of fairness, development of explanations from ML models, robustness analysis, etc.) to provide support for different stakeholders along the ML lifecycle;
- However, there has been less work done for the "middle-level", where some extent of generalisation across use-cases and practical guidance and best practices for users is both possible. The Responsible ML tool elaborated by *AWS* helps fill this "middle-level" gap;
- The intention was to provide support for critical thinking for a wide variety of case studies in the form of recommendations on how to consider important aspects of a responsible ML system. Since responsible ML is a rapidly developing field, the recommendations formulated by the tool will also evolve over time and the tool will be updated to reflect feedback and ongoing scientific developments;
- *AWS* also offers additional services and tools, educational and scientific resources, and Professional Services where ML experts offer consultancy services to help aid responsible AI case studies. This existing larger system along with existing third-party tools can be leveraged and should also be taken into consideration when applying the *AWS*'s Responsible ML tool.

2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it embedded/integrated in the AI product development processes (of the user of the tool)? How does it support the continuous monitoring of an AI product during usage?

AWS Customers and developers of ML systems, in general, are increasingly putting effort to voluntarily make their ML systems more responsible. Customers have expressed interest in *AWS*'s views on this topic since responsible ML

Table 2. AWS ("specialized") tools and resources for responsible ML systems

AWS tools and resources for responsible ML systems
– **AWS SageMaker Clarify** – **AWS Augmented AI** – **AWS SageMaker Model Monitor** – **AWS SageMaker Data Wrangler** – **Training and Professional Services** – **Research, Innovation, and External Collaboration**

has become core to many of their organisations and being perceived as "unfair" is not an option for businesses. To support these needs the tool provides recommendations and examples that can be used across three major phases of the ML lifecycle: (1) design and development; (2) deployment; and (3) operation.

To reach the largest possible audience, the tool will be available both for customers of *AWS* and for the larger public looking for guidance on responsible ML systems. Therefore, it can have added value for all users/developers/ deployers/etc. of AI and support a systematic reflection at any stage of the ML lifecycle, although the tool should ideally be consulted even before the ML development begins.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

Several internal *AWS* teams were involved in the development of the tool with different professional profiles including cross-section of business and technical leaders, public relations, public policy, and legal departments. For instance, cross-functional contributors were valuable in making this complex topic easily understandable for a larger audience. They contributed to finding the right balance between precise technical wording and framing that is more understandable by a larger public. *"We needed to find the sweet spot between how we go about communicating complex questions in a simple way that does not dilute the essence of what you are trying to say"*. There have been also substantial exchanges with *AWS* business owners to provide feedback during the development of the tool: *"we also benefited from the view of the business owners who could bring in the type of questions that their customers would ask".*

While the tool tries to help the organisations that are looking for support in making their AI/ML solutions more responsible, it also intends to bring organisations' internal stakeholders even without any deep data science knowledge to a minimum common level of understanding concerning these solutions.

Table 3. Three phases of the ML lifecycle addressed by the Responsible ML tool

Phase 1: Design and development	Phase 2: Deployment	Phase 3: Operation
– **Evaluating Use Cases** – **ML Capabilities and Limitations** – **Building and Training Diverse Teams** – **Be Mindful of Overall Impact** – **Training and Testing Data** – **Bias** – **Explainability of ML systems** – **Auditability** – **Legal Compliance**	– **Education, Documentation and Training** – **Confidence Levels and Human Review** – **Testing and re-training** – **Notice and Accessibility** – **Safety, Security, and Robustness** – **Legal Compliance**	– **Provide and Use Feedback Mechanisms** – **Continuous Improvement and Validation** – **Ongoing Education**

This facilitates communication between the technical, non-technical, and leadership teams and makes the responsible ML theme more "digestible" at all levels of an organisation. However, independent from the specific use case, the tool can also help facilitate buy-in on behalf of top-level management by providing a framework for discussion in non-technical terms, which can be helpful for the efficient adoption of responsible use principles in organisations. External to an organization, the tool can also provide information and assessment on, for instance, how a vendor is following the process to build responsible ML systems.

4. What are suggestions to consider while creating a Responsible ML tool?

Based on the lessons learned from the development process of the *AWS* Responsible ML tool, the following elements are observations and suggestions to be potentially considered when building similar tools:

- **The aim and the scope of a Responsible ML tool can be better understood** through placing it within the context of other existing tools and frameworks. The scope of the instruments within this system can be different in terms of potential for generalisation across use-cases and practical guidance;
- **Considering how people understand AI and tools is key.** The involvement of non-technical experts in the development of the tool was critical in the case of this *AWS* tool. This facilitates the development of comprehensible tools – as comprehensibility of a tool can be important for the adoption of that tool;
- **Continuous improvement and updating of the tool** are key as ML technology and related use cases are constantly evolving. This needs to also be reflected by the tools used along the development process of ML products. Relying on feedback both from stakeholders outside (e.g., clients using the tool) and inside (e.g., internal users of the tool) the organisation is the optimal way for continuous improvement;
- **Responsible ML systems tools can also have additional positive impacts**, such as increasing organisational engagement, helping communication about ML systems across the organisation and supporting ML evangelisation;
- **The tool was developed through a process of alignment** ("accept it, digest it, improve it"), which is key especially in a large company. This is a complex exercise and needs buy-in and a significant investment of time and effort at all levels of the organisation;
- **Standards (e.g., published by IEEE, ISO, etc.) will probably become increasingly important** in guiding for ML use-cases compared to the "checklist type of tools". These are reaching an important level of maturity with some level of abstraction but a relatively precise focus on specific ML use-cases.

Fairness Compass of AXA

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

During the past years, different types of biases related to AI systems have been demonstrated, and professionals developing AI solutions often face the complex issue of "how to implement fair AI". Today, the answer to this question is mostly provided by data scientists who fine-tune data and algorithms to satisfy some existing definitions and measures of fairness without necessarily considering the general context the system will operate in. However, "sustainable solutions for fairer AI must go beyond technical methods and explain what the implemented fairness objective stands for and why this choice was considered most suitable for the given scenario".[6]

The responsible AI team of *AXA*, a French global insurance company, has developed the Fairness Compass tool with multiple aims. First, the tool intends to provide a clear and transparent process by supporting the decision-making concerning the fairness objectives of a given AI solution. Second, the tool aims to document the decision-making process about the choice of fairness objectives and to provide clarity for all stakeholders, who may eventually also question the AI project about how any specific fairness metric has been chosen. Finally, the aim of the Fairness Compass is also to include policy professionals upfront in the choices of AI fairness objectives and metrics, not only the technical teams (e.g., data scientists), as policy professionals have a broader understanding of the general context and can provide a clear roadmap for the technical team in relation to the choices and implementation of AI fairness objectives.

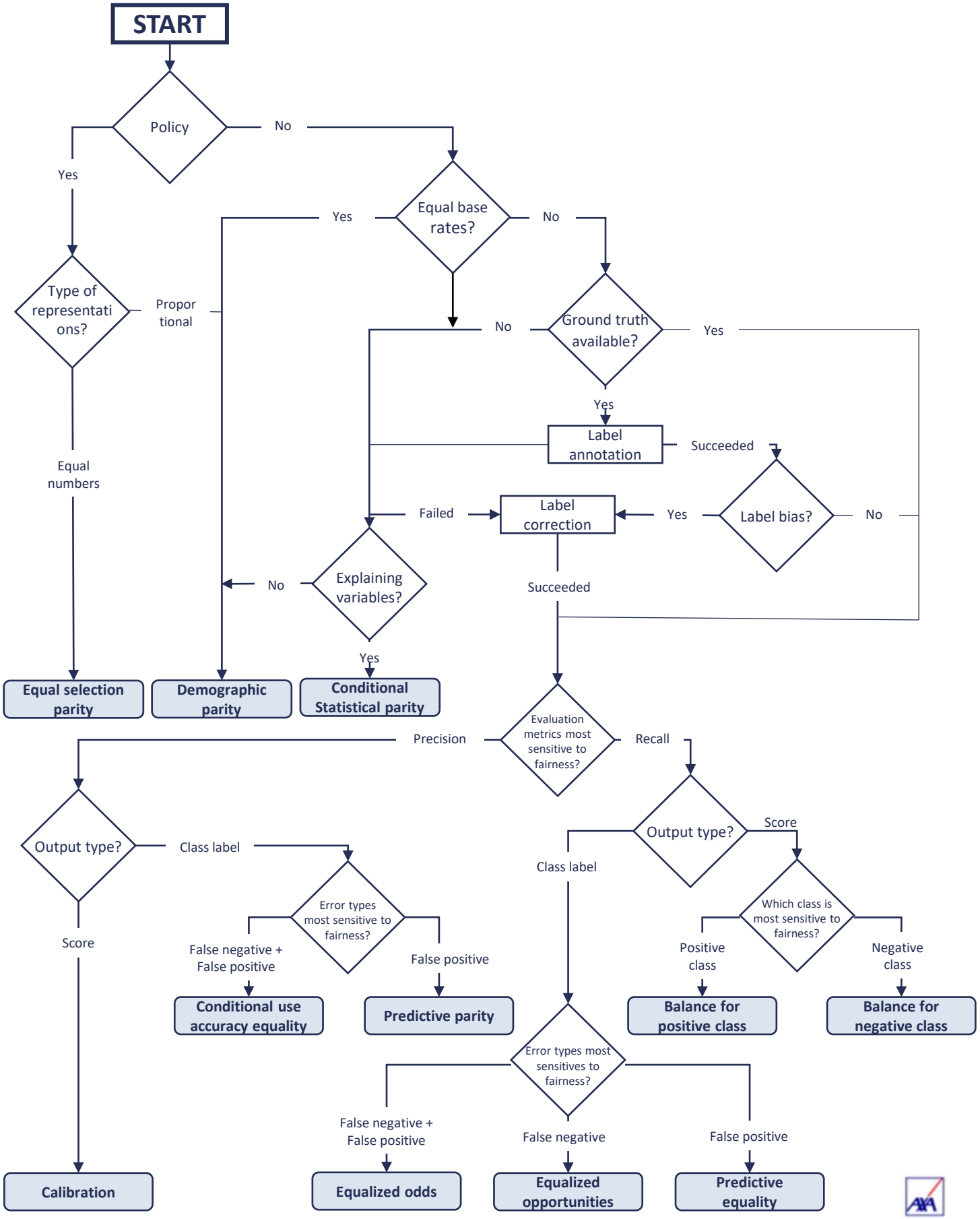
2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it embedded/integrated in the AI product development processes (of the user of the tool)? How does it support the continuous monitoring of an AI product during usage?

In practice, the Fairness Compass is a comprehensive decision tree in the form of an interactive graph containing a set of contextual questions such as about "the nature of data, beliefs in its correctness, fairness policies, and on specificity versus sensitivity of the model".[8] The tool is designed with great flexibility since it allows its users to modify the graph to better adjust the decision tree to their specific context.

Identifying the fairness objective as a first step can be as important as "debiasing the algorithm". After all, debiasing AI based on a less appropriate fairness metrics may likely not be satisfactory. The tool structures and prioritizes the complex landscape of fairness definitions for the user, so the very first decision is about how the given AI solution relates to the policy, legal, and regulatory context. As Boris Ruf, Lead Expert in Algorithmic Fairness at *AXA* explained: "For example, people may expect different types of fairness. So, we need to provide more information on which fairness metric the application is trying to achieve and why". Another important aspect of the tool is that it documents how the decision has been taken. This can be referred to later, also "enforcing" commitment on behalf of the decision-makers. Documentation is key when it comes to implementing fair AI – and generally AI systems.

The decision nodes along the graph lead to different fairness metrics, although different interpretations are possible even for the same use case.

Figure 1. Fairness Compass decision tree[7]



Based on B. Ruf and M. Detyniecki, "Towards the Right Kind of Fairness in AI", ECML/PKDD 2021 (Industry Track)

Therefore, debates and involvement of people with diverse views is important. The tool is used during the design phase of the AI lifecycle, and depending on the outcome, different bias mitigation approaches are possible.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and their responsibilities? Does the tool require specific skills or training? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

The result of the tool is a documented decision (and decision process) about the AI fairness objectives which are provided as input to the development phase. It is recommended that the policy-level decision-makers using the tool have some level of knowledge about data science and the concept of fairness in relation to AI systems. Moreover, as some questions are very technical, for example about data quality, output type and precision/recall or other statistical metrics, the active involvement of roles such as lawyers, ethics advisors, data scientists or data engineers is important. Currently, the responsible AI team at *AXA* is supporting its policy-level experts internally about how the tool should be used.

For the moment, the Fairness Compass' intentional use is internal to the organisation developing the AI solutions, but eventually for transparency reasons information about the use of the tool should also be available to customers outside the organisation.

4. Example use cases

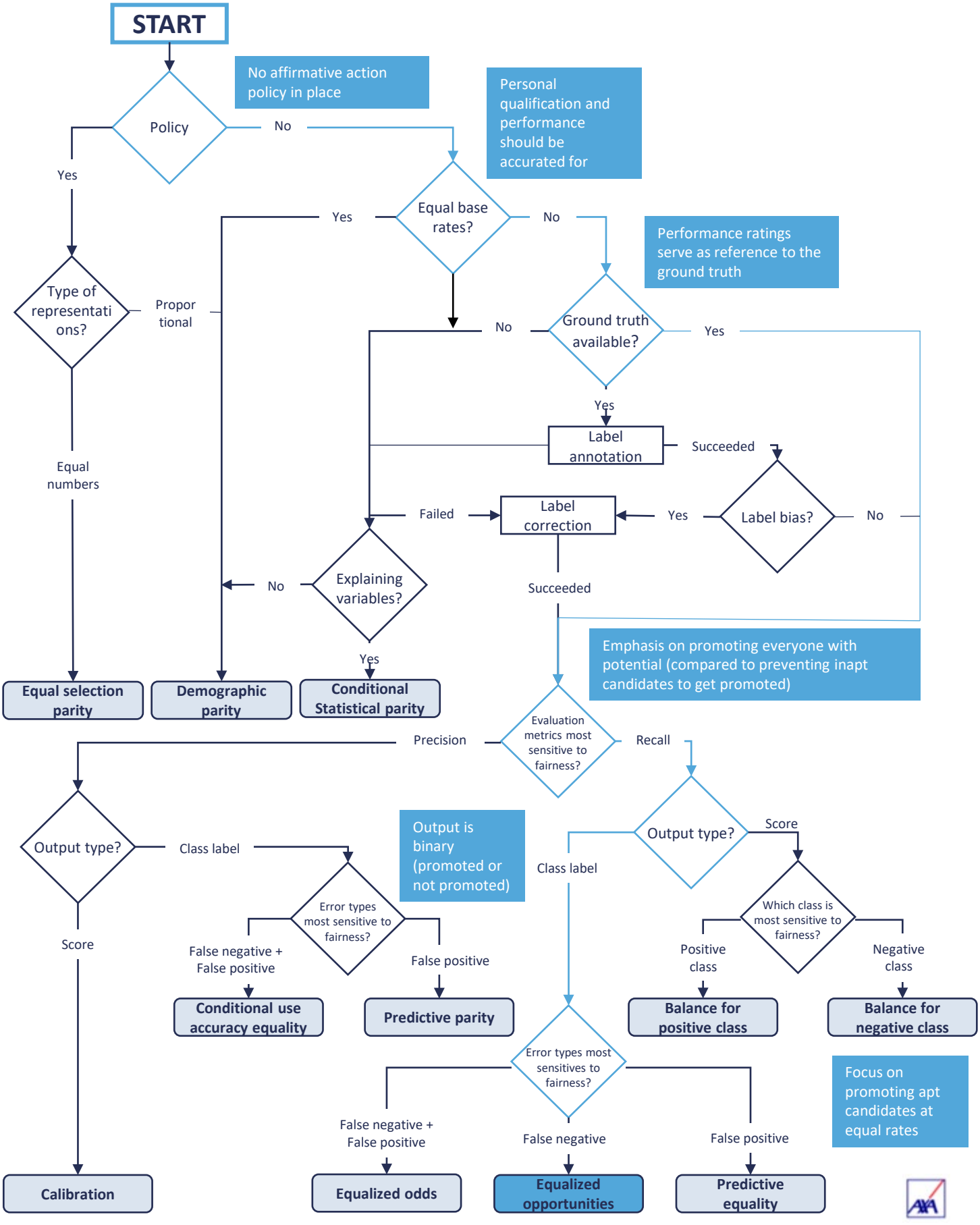
A sample scenario in the context of human resource management illustrates the functioning of the Fairness Compass. The sensitive subgroups considered in this example are men and women. The question of interest is which definition of fairness would be most appropriate when it comes to assessing fairness in employee promotion decisions. Obviously, this is just a fictional thought experiment, and depending on the context, other answers with different results may apply. The purpose of the Fairness Compass is to support well informed decision making based on the defined requirements for a given scenario.

In Figure 2, the Fairness Compass is represented as a decision tree with three different types of nodes: The diamonds symbolize decision points; the white boxes stand for actions and the grey boxes with round corners are the fairness definitions. The arrows which connect the nodes represent the possible choices.

After starting the process, the first question is about existing policies which may influence the decision. Fairness objectives can go beyond equal treatment of different groups or similar individuals. If the target is to bridge prevailing inequalities by boosting underprivileged groups, affirmative actions or quotas can be valid measures. Such a goal may stem from law, regulation, or internal organizational guidelines. This approach rules out any possible causality between the sensitive attribute and the outcome. If the data tells a different story in terms of varying base rates across the subgroups, this is a strong commitment which leads to subordinating the algorithm's accuracy to the policy's overarching goal. For example, many universities aim to improve diversity by accepting more students from disadvantaged backgrounds. Such admission policies acknowledge an equally high academic potential of students from sensitive subgroups and considers their possibly lower level of education rather as an injustice in society than as a personal shortcoming.

For the sample scenario, the stakeholders may conclude that no such affirmative action policy is in place for promotion decisions. Therefore, they may choose "No" and document the reasoning behind their choice. This procedure is repeated question after question until a leaf node is reached which contains the recommended fairness definition. In this case, the outcome is "Equalized opportunities", a concept which ensures that the probabilities of being correctly classified are the same for everyone.

Figure 2. A fictional example of the use of the Fairness Compass: Promotion decisions



5. What are the current limitations of the tool? What are the possible recommendations for regulators?

Best practices for creating the tool and its limitations

The Fairness Compass can be used as-is with general examples from all domains. However, while such examples can provide important guidance, it is also possible to customize the Fairness Compass and provide domain-specific examples. The use of clear and relevant examples while developing the flowcharts of the Fairness Compass is one best practice identified by the *AXA* team. The tool has also been designed in such a way as to allow such customisation.

However, a current limitation of the Fairness Compass is that it is only adapted to classification problems, and also does not help with the further technical implementation of the fairness metrics. There is ongoing research to adapt the tool to more complex regression (and other) problems, too. Furthermore, the development of a dedicated library of (technical) methods related to each potential outcome of the tool's decision tree/graph. This will allow to statistically achieve the desired fairness metrics (e.g., adversarial approaches, weighing approaches, etc.), which can be later implemented by data scientists.

Feedback to regulators

The following observations might provide valuable input to regulators both at an international and national levels:

- **The industry needs to have clear guidance on how to solve the problem of AI fairness.** Fairness is a complex issue, and it can have multiple definitions depending on the context. Consequently, just requiring "fairness" from a regulatory level is too general, and needs more practical guidance and tools;
- **Targeting the wrong metrics to achieve fairness in AI can lead to a result that is even worse[9] compared to a "biased" outcome according to some such metrics.** The tool tries to avoid this since it helps policymakers, developers, and possibly deployers and business users to define the type of fairness the (end) users are expecting;
- **The process of fairness metric selection and documentation is an important added value of the tool that also provides information for regulators.** The commitment of the organization's policymaking team is also crucial, as this group has joint responsibility (together with others, e.g., the management and technical AI teams) about implementing fair AI systems;
- **There are often limitations concerning the collection and sharing of sensitive attributes** (e.g., gender, age, etc.) which can make the creation of fair AI solutions more difficult. Even more so, algorithms can "learn" biases through different proxies (e.g., insurance-related data embed possible biases that the driver of a red *Cinquecento* may be more likely to be female – a frequent perception). However, the correction of such biases can be improved through accessing sensitive data. A possible way to overcome this problem would be to design architectures where companies can use sensitive data without allowing public access to those. For instance, some third-party entities could have the legal authorisation to both collect sensitive information and evaluate AI models developed by companies and organizations in each industry.

Meta's "Why Am I Seeing This" (WAIST) Explainability Tool for Facebook

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

Being the largest and most visible social media platform globally (based on Statista, the number of *Facebook* users has grown from around 1 billion in 2011 to around 2.85 billion by July 2021), *Meta* has attracted a lot of media attention as well as criticism about how content is provided to its users.[10] For many reasons, ranging for example from improving users' satisfaction, trust, and engagement, to improving *Meta*'s services based on users' feedback, or to managing potential reputation risks, it is important for *Meta*'s to provide explanations for the choice of recommended content, connections or advertisements displayed to users. To this purpose, *Meta* has developed several tools, processes and practices to increase transparency and explainability about the specific content served on its platform, also in consistency with guiding principles, such as the OECD AI Principles of "Transparency & Explainability" and "Accountability".

Meta takes a broad approach to transparency in this area. First, it focuses on two different forms of transparency: Process-based transparency tools and outcome-based ones.[11] Process-based tools focus on how decisions, such as content selection, are made in general, possibly including information about the inputs used. An example is one component of *Facebook*'s "Favorites" feed tool[12] that shows which friends are considered the most meaningful to a given user – the tool also gives the user the possibility to change this list and tell *Facebook* directly which friends are their "Favorites".

Outcome-based tools explain more the specific outputs of an algorithm, e.g., which outputs were reached or why particular outcomes were reached. The "*Why Am I Seeing This*?" (WAIST) tool on *Facebook*, the focus of this case study, can principally be considered an outcome-based explainability tool, but *Meta* has others, such as its recently launched Widely Viewed Content Report.[13] Second, within those types of transparency, *Meta* uses different mechanisms to show and explain relevant information. Mainly, *Meta* considers three approaches: (a) tools-based transparency – transparency that is provided through the user-interface, like the Favorites and WAIST tools; (b) prose-based transparency – transparency that is provided through written information or explanations, such as blog posts in the *Meta* Newsroom[14] about ranking changes[15] and how ranking works[16] or the policies provided in *Meta*'s Transparency Center[17], like its Community Standards[18]; (c) data-based transparency – quantitative information about the content that appears on its platform, such as its quarterly Community Standards Enforcement Report[19].

WAIST is one of the tools *Meta* developed to increase transparency and to provide an understandable explanation to users on why they are provided with a specific content on *Facebook*. It is an example of tools-based transparency that includes information about process and outcomes. The tool also allows and facilitates a set of actions users can take to influence the selection of the content provided to them going forward. *Meta* uses an AI based multiple layer system to select from the inventory of posts (e.g., posts from friends or connected Groups) the specific content that will be provided for each user. The ranking algorithms focus mainly on identifying the couple of hundred valuable pieces of content out of the likely thousands of them in a person's inventory, which is a more complex task then further selecting the top posts out of these few hundred identified (see Figure 3). Therefore, the content shown in a user's News Feed does not

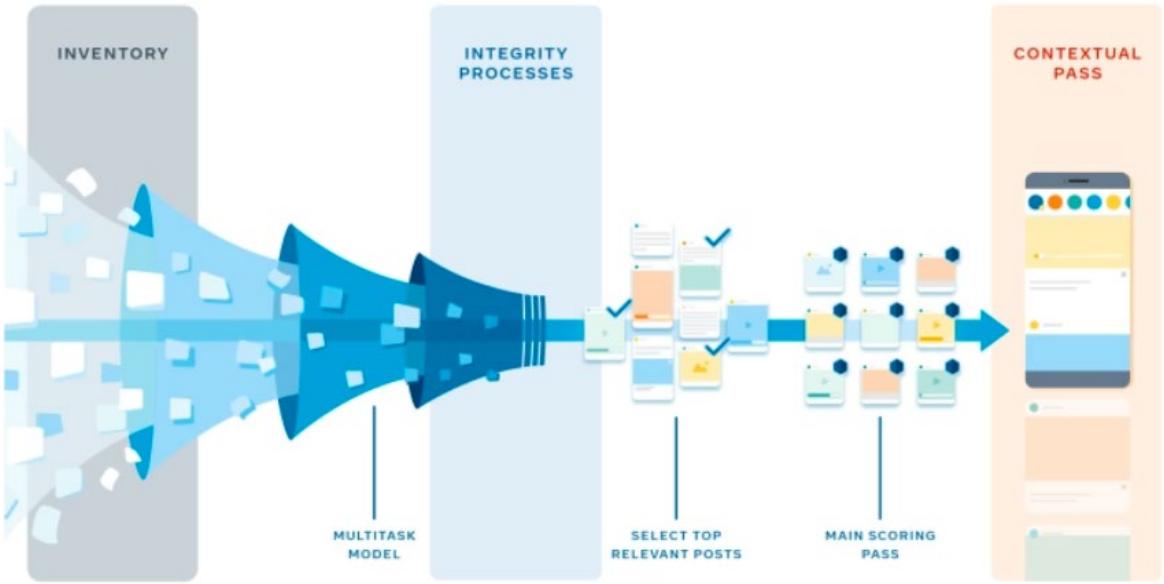
necessarily appear in chronological order but can be defined through a content ranking process and algorithms that are constantly refined over time.[20]

2. How does the tool work?

Although social media platforms – including *Meta* – have faced criticism for supposedly focusing on user engagement, *Meta* claims that its content ranking is optimised to train a series of models to provide the most valuable, relevant, or meaningful content to users based on algorithmic predictions. This is achieved using a value score that reflects a number of different predictions that, taken together, are designed to approximate value. For example, the ranking process assesses the probability a user will like the post, think the post was worth their time, comment on the post, or hide the post, and it also assesses the likelihood that the post will be clickbait or include a highly exaggerated health claim. Some of those assessments act as positive inputs into the value score and others as negative inputs. Many of the negative inputs encompass types of content and behavior *Meta* thinks is problematic in some way but that does not rise to the level of removal. *Meta* recently published the list of those negative inputs, what it calls its Content Distribution Guidelines[22], as another important form of prose-based transparency. The final value score for any given post for any given user accounts for all the positive and negative inputs, the full range of predictions.

Personalized predictions of the probabilities for the possible actions detailed above are made based on thousands of signals. For each specific content, the WAIST tool's aim is to give an easily understandable explanation to the user in relation to the final outcome (the selected content). The tool mainly provides information on the most important three signals in the ranking process, explaining how they factored into the ranking of particular posts to a particular user. However, WAIST does not provide information about how all signals are used, the relevant weights/importance of the signals that are not included. The tool is limited to showing how the three most important signals influence the ranking of particular posts because of the outsized influence played by those signals and to balance explainability with transparency, i.e., to ensure that the information provided by the tool would be easily understood – and meaningful – across *Facebook*'s broad user-base. A key insight, based also on user research *Meta* conducted, is that transparency (e.g., sharing information about all, or many, signals used by the *Facebook* algorithm) does not always enhance explainability – for example as users are

Figure 3.[21]



not necessarily able to understand very complex "explanations". Further, people sometimes find dense explanations to be a bad experience, because they do not come to the platform to be bombarded with educational explanations about AI, they just want to use the platform to connect. Therefore, *Meta* identifies the most important factors driving the algorithm's output, which are also of interest and are easily understood by the users and uses only these factors to provide explanations.

The tool itself is accessible in a drop-down menu in the right-hand corner of a post on *Facebook*. It provides explanations and possible action items for the users (see Figure 3), relating to:

- *Why a specific post is shown to the user*. The tool will for instance give the explanation that the post is shown because it is from a friend, a Group one joined or a Page one follows;
- *What signals generally have the largest influence over the order of posts*. Three signals are used for the explanations: (a) how often one interacts with posts from other users, Pages or Groups; (b) how often one interacts with a specific type of post, for instance, videos, photos or links; and (c) the popularity of the posts shared by the users, Pages and Groups one follows. These are generally the three most influential signals in the ranking process.[23] The tool only endeavours to show how those signals factored into the ranking of a piece of content – it does not show the other thousands of less important signals that could have influenced a post's ranking. This helps *Meta* further its goal of explainability. It aims to provide a comprehensible but meaningful amount of information about the ranking outcome, rather than overwhelm them with loads of information that will not be useful to them. This strikes a balance between transparency (e.g., in the extreme case providing access to the algorithm itself, including the thousands of signals it uses) and explainability (e.g., considering human factors that affect how understandable but also action oriented an explanation is);
- *Tools that enable the user to take specific actions* given the explanations provided, such as to unfollow a friend or Group, hide advertisements, or manage preferences or privacy settings. This is considered a critical component as it provides some control to the user and enables actions given the provided explanations. Based on focus groups and other customer insights initiatives, *Meta* believes that explainability and control are related and therefore there is benefit in thinking about them together. Explainability is important to understand why algorithms make the decisions they make, but it is also important to exercise some control by being able to take actions given the understanding of how these decisions are reached. This includes enabling users to change and influence how the system produces its decisions going forward, or to reduce and mitigate the effects of these decisions.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

A critical success factor for the development of WAIST has been to start from "who are the key users of transparency" and then spend time to understand their needs. For instance, the company has been using focus groups as well as surveys (e.g., asking users questions such as "did you like the post you saw?") to understand what matters for its users. Understanding the audience's needs is considered a key best practice when developing explainability tools. Perspectives from experts, such as academics, civic society, and sophisticated users, have also been gathered during development.

WAIST is relatively easy to use, not requiring any specific training – although *Meta* provides information online on how to use the tool. A key design choice that makes the usage of the tool simple is the *consistency of design* across all applications (e.g., explanations for advertisements, recommended content or recommended connections – see below), even if the algorithms used to generate the explanations may differ across applications. This choice also makes it easier to build the tool across different applications. Although WAIST is potentially usable by other companies, the tool has only been implemented at *Facebook* for the moment. However, the principles followed to develop the tool as well as the design choices made can be used by other companies to implement similar tools.

The WAIST tool is oriented towards customers, but as noted above, *Meta* has other transparency tools and practices. Different forms of transparency can better serve different types of audiences. Some provide even more information about how content selection and other functionalities work. For example, *Meta*'s CrowdTangle tool[24], which provides outcome-based transparency about posts from public *Facebook* Pages, e.g., which posts have received the most interactions, which can be filtered within specified boundaries, e.g., time, topic, geography, is mainly geared towards researchers, publishers, creators, and journalists and not towards everyday consumers.

4. Example use cases

Meta classifies content on *Facebook* into three different groups: advertisements, connected content and recommended content. WAIST has been implemented for all three content categories.

WAIST was originally developed for advertisement content in 2014 to increase transparency for users around why they are seeing specific ads on *Facebook*. WAIST displays selections made by advertisers to define target audiences that match users by age, gender, location, and one or more audience selections.

The tool was later introduced also for connected content. Connected content relates to the users' direct contacts, the Pages they follow or the Groups they are members of. By estimating the "importance" of, say, a friend's post, this type of content is also ranked before being shown to users. There is also a control element to this type of content allowing for instance to unfollow friends, etc.

Finally, WAIST has also been developed for recommended content. This is the type of content that has been recommended to users by *Facebook*'s algorithms based on their previous activities or preferences, and it does not relate to their connected content. The most complex part of the algorithmic selection process is to narrow down the recommended content available on *Facebook* from billions of posts to approximately 200 chosen ones and not to select the specific content to serve out of the 200 posts considered. WAIST again combines explainability with control, allowing users to unfollow friends or to opt-out from the group to which *Facebook* is classifying them, for instance.

5. What are suggestions to be considered while creating a similar tool?

Some key decisions had to be taken during the development of the tool since it is difficult to find consensus around "what is a good explanation":

- **Transparency and control are strongly related, and it depends on the context whether both are necessary.** While building WAIST, the preferences of the target audience were gradually understood and it became clear that in addition to some transparency, "giving control" is also critical. This meant that explanations needed to be also action-oriented to provide the user options to take actions (e.g., delete themselves from a given audience that advertisers were targeting or remove links from Friends or Groups, etc.);
- **There is no "one methodology fits all solution" for explanations.** All the example of transparency case studies discussed above are different.

In the initial phase of WAIST's development, the expectation was that the exercise would be the same for each case. This was a false assumption, and a different methodology was finally developed for each case;

- **On the other hand, consistent and standardised design is important**. A similar design was used across all applications of WAIST, e.g., advertisement, connected content, recommended content. This helped users to better understand the tool and made it easier for them to adopt it;
- **Transparency is not equivalent with good explanations. Explanations should include signals/factors used by the algorithms that are likely to be the most relevant and understandable to people.** That is, tools-based transparency in particular should focus on factors that are especially important in the ranking process and investment should be made to explain signals and factors in a way people can understand. Being transparent about how ranking algorithms operate but in a way that is not easily understood is not useful for most external audiences. Finding a balance between full transparency and good user-centric explanations – which are also action oriented – is key. User feedback is key to achieve this;
- **The explanations may need to be approximate.** It is often not useful to explain things to people in the precise, scientific terms that are used to program an algorithm. Rather, especially when it comes to explaining a process or the "why" of an outcome, providing explanations that closely approximate the complex functioning of the algorithm best serves the twin goals of transparency and explainability in a balanced way;
- **Feedback from users and other specific audiences are crucial during development.** The perspective of the different user groups, academics, civil society can only be understood through constant consultation and feedback. *Meta* has also exchanges with experts to fully understand the possible societal and political implications of the transparency tools, as well as the philosophical benefits of transparency (e.g., what do the tools mean for polarization, civic society, etc.);
- **Transparency can take different forms.** It is important to decide during the development process what form transparency will take. Outcome-based transparency tools provide explanation about the specific outputs of an algorithm while process–based tools focus on how decisions, such as content selection, are made in general, possibly including information about the inputs used. Within those types of transparency *Meta* considers three approaches: (a) tools-based transparency – transparency that is provided through the user-interface; (b) prose-based transparency – transparency that is provided through written information or explanations and (c) data–based transparency – quantitative information about the content that appears on its platform is used by *Meta;*
- **There are different types of users, and transparency tools need to be useful and accessible for the entire community.** Explanations need to be provided in ways that are both understandable and engage users to continue expecting and requesting explainability and transparency:
    - **Language is crucial both for transparency and explainability**. Having the tool available in as many languages as possible is important;
    - **Regional differences have to be taken into consideration**. There has to be interaction with all subgroups of users and the main themes, usages and needs have to be understood and addressed in all geographies;
    - **Digital literacy has to be assessed** and the transparency tool needs to be developed based on this assessment. Users need to be met at the level where they are in their digital literacy journey.

Based on *Meta*'s experience developing WAIST for *Facebook* and other explainability and transparency tools and practices, some recommendations for regulators can be made, specifically:

- **Regulators need to allow different approaches to achieve transparency and explainability.** Regulators increasingly consider transparency and explainability as regulatory requirements for AI systems. However, for transparency and explainability to be the most effective, regulators need to work with industry experts to come up with guidance on how to satisfy transparency, increase explainability, and offer control to users. Striking a balance between these three goals may be challenging without some flexibility to use different approaches;

  Providing information about transparency activities can be a good solution. One option can be to let companies demonstrate what they do for transparency through documenting transparency practices and processes, as *Meta* has done with WAIST and other tools;

- **Targeted transparency and digital literacy.** Regulators need to formulate their guidance and positions on transparency according to the specific groups that are receiving those explanations since user groups can be very different (e.g., from a digital literacy point of view). They may also need to couple transparency requirements with digital literacy programs.

IBM FactSheets

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

IBM AI FactSheets 360 is a tool that aims to capture essential information through a common and standardized set of attributes about how an AI system was developed, tested, and is intended to be used. Its goal is to increase transparency of how an AI system was built and support accountability. These model "facts" can include information about fairness, explainability, privacy, adversarial attacks, uncertainty quantification, etc., all of which can be computed by other "360" tools that *IBM* has produced (AI Fairness 360, AI Explainability 360, AI Privacy 360, Adversarial Robustness 360, Uncertainty Quantification 360).[25] Hence the tool being most relevant to the OECD AI Principles of "Transparency & Explainability" and "Accountability". It is also relevant to all AI system lifecycle stages, and to both "high-risk" and "limited-risk" categories of the proposed EU AI Regulation. It is flexible and can capture whatever information is important, such as what is specified in principles, regulations, or local best practices. In relation to the draft OECD AI Classification framework, the FactSheets methodology is flexible enough to capture information from all four dimensions. The main focus has been on "2. Data and Input" and "3. AI Model", though it also can address key elements of "1. Context" and "4. Task & Output".

From a value creation point of view FactSheets can help companies improve the effectiveness of their current AI development, management, and governance processes. For example, it can be used to enforce enterprise governance policies, by only allowing the AI development, implementation, or usage process to advance to the next stage if a fact condition is met. This can provide a required level of control, for example to minimize risk and ensure compliance with regulations.[26] The tool also enables AI stakeholders such as risk managers, engineers, or customers to determine when it is appropriate to use the AI system, potentially improve it, better understand its ethical and legal concerns, and support AI governance.[27]

2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it integrated into the AI product's development process? How does it support the continuous monitoring of an AI product during usage?

Building and operating an AI system is a complex task and the result of the work of multiple professionals with different backgrounds and roles throughout the AI lifecycle. This means that each of these professionals impacts the final end-product in ways that others may not be familiar with, but also that they can produce specific information about their work which can be included into the AI system's documentation. This information is collected throughout the AI lifecycle and is presented in the form of FactSheets.

Table 4. AI lifecycle roles and related AI facts[28]

	Role in the AI lifecycle	Example AI model facts generated
Business owner	Defines business roles and requirements	Facts about model purpose and governance
Data scientist	Uses data to train models to meet requirements	Facts about data transformation, features and performance
Validator	Uses business goals, regulations and best practices to test models	Facts about fairness, privacy, functionality and verification
AI operations engineer	Deploys and monitors models in running services	Facts about performance drift, learning, and monitoring

FactSheets are, in practice, the collection of answers to a well defined – but also customizable – set of questions. These can concern for instance the general purpose of the model, the characteristics of the datasets used to train and test the model, the design and trade-off choices made during the model creation and deployment, the performance and possible biases of the model, or the conditions under which the model may or may not operate appropriately – hence also supporting its monitoring during usage. In general, developing these questions is done through an iterative process, mainly by interviewing the persons involved in the AI lifecycle (Table 4), and is orchestrated by a designated "FactSheet team". Table 5 shows a possible methodology for such a process. Some of the information gathering, particularly related to the performance of the AI system (e.g., accuracy, fairness metrics, robustness, etc.) can be automated. The documentation of these performance metrics, and potentially their continuous update during the usage of the system, can help the organizations using the AI systems to ensure the fairness, robustness, and safety of these systems.

A level of customization of the questions to be answered and the information documented is needed, as not all AI systems are similar and not all stakeholders have the same needs in terms of information. However, customization can be facilitated following a process, for example by first understanding the needs of the AI system end users and their preferences and requirements for the information to be captured in the FactSheets.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

FactSheets can provide benefits for *stakeholders* both inside and outside the AI lifecycle and both inside and outside the organisation developing the AI system (e.g., customers). The *first level of users* are those who are part of the AI lifecycle (Table 4) and are also actively contributing to the AI system development. The facts produced by these "fact producers" become the input for "fact consumers" during the AI lifecycle, as for instance the information from a data scientist about the performance of the model will be an essential input for the AI operations engineer.

Table 5. AI lifecycle roles and related AI facts[29]

	Steps for constructing AI Factsheets	Responsible party
1	Gather the information needs of potential FactSheet consumers	FactSheets Team (with potential consumers)
2	Gather the kinds of information FactSheet producers might generate	FactSheets Team (with potential producers)
3	Define the topics and questions to be included in FactSheets	FactSheets Team
4	Informally assess FactSheet Template by trying to fill it in	FactSheets Team
5	Populate a FactSheet Template with actual facts	Business Owner, Data Scientist, Model Validator, AI Operations Engineer (and others as defined within an organization's AI lifecycle)
6	Assess FactSheet quality with those who will be consuming FactSheets in production	Business Owner, Data Scientist, Model Validator, AI Operations Engineer (and others as defined within an organization's AI lifecycle)
7	Evolve existing templates and create new ones	FactSheets Team (and others as appropriate)

A *second level of users* are those who are outside the AI lifecycle while being part of the organisation that has itself built the system, but did not necessarily contribute to constructing it, such as the risk department, the ethics board or the legal department.

The *third type of users* are those outside the organization that developed the AI system. These can be for instance a *regulator*, who may audit the AI development processes as well as the functioning of an AI system, or customers/users of the AI system. Users can be both those operating the AI system and others affected by it. For example, in the case of an AI medical device, a typical *end user* would be a medical doctor using the AI tool to detect skin cancer and would be interested in its level of accuracy. An *affected user* would be a patient who would like to have a general understanding of how the AI system works, of its possible biases, or its robustness and reliability.

4. Example use cases

There are several examples of use cases[30] developed by *IBM* on how Factsheets can be built in practice. One example relates to the *audio classifier model*[31], which classifies an input audio clip into five different classes it detects. A first version of the related FactSheet was developed based on interviews with the data scientists who built the model at *IBM*. This helped identify the possible information about the model that an AI developer/customer who wishes to use it would need. This first version was then further refined through several iterations with field experts and possible users. The whole process represented approximately one week of effort.

Another example is a *mortgage evaluator governance model*[32] that predicts mortgage approval. The FactSheet for this model was the result of both an iterative interview-based process as noted above, regarding for example free form questions such as the purpose of the model or the required tests for the model ("model policy"), as well as an automated process developed by *IBM*. The latter was used to capture information about the work of the data scientist (such as possible data manipulations, definitions of new features, or tests performed) while building the model. Ultimately such an automated process can greatly facilitate the development of FactSheets in accordance with corporate level "model policies" (e.g., information requirements regarding the accuracy, fairness, explainability, or robustness of the AI system).

5. What are suggestions to consider while creating a Factsheet?[33]

Data scientists, internal and external to *IBM*, were both involved in testing the FactSheets. In general, AI system developers considered FactSheets highly valuable because documentation helps collaboration, model reuse, maintenance, and improvement. As a result, the following points are suggestions to be considered during the development process of FactSheets:

- Details of the model development have to be recorded during the process as attempting to reconstruct them later is highly time and resource consuming;
- There is no "one size fits all", and customization of FactSheets is needed;
- Maximising the involvement of stakeholders in creating FactSheets is important, in order to obtain a well-balanced feedback and avoid a mostly technology-driven process;
- Data scientists have to be aware of the benefits of using FactSheets for their work and for others, as they are key stakeholders in the process;
- It is important to find the right balance between providing enough information about a model and not revealing information that is proprietary or harms business interests;

  Proper attention has to be given to different type of biases in the model or the dataset, as AI developers can be unfamiliar with the notion of bias;
- Automation of parts of FactSheets can highly facilitate its use in production.

Challenges exist in all seven steps involved in applying the tool's methodology[34], if not performed correctly, such as: not properly understanding who are the consumers of the FactSheet; what kind of information they need to see about an AI model; the difficulty of capturing relevant information at the time it is created (and the time/effort needed to do that retrospectively); the lack of a one-size-fits-all approach etc.

Responsible AI Governance, Transparency notes and Error analysis tools at Microsoft

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

Microsoft is a leading multinational technology corporation, headquartered in Redmond, Washington (USA). Over the past several years, the company has made significant progress in developing and using artificial intelligence (AI) technology. This has been made in line with ethical guidelines to ensure that potential AI-related risks are properly anticipated and mitigated while maximising their benefits to businesses and society.

Microsoft has based its AI policy on six responsible AI principles (Table 6) and is translating these into its practices through an effort led by three internal teams. The Aether Committee advises the leadership team on the challenges and opportunities presented by AI innovations, the Office of Responsible AI puts principles into practice by setting the company-wide policy for responsible AI through the implementation of internal governance and public policy work. Finally, Responsible AI Strategy in Engineering (RAISE) is an engineering team built to enable the implementation of responsible AI policies and processes across the engineering groups.

To support its AI ethics goals, *Microsoft*'s AI governance approach follows a "hub-and-spoke" model that helps the company to integrate privacy, security, accessibility, and responsibility into its products and services: "*Our approach is operationalising our AI principles by establishing an internal governance structure, and setting the rules the organisation needs to follow. This is in addition to developing the practices that help our teams follow the policy requirements and work with a human-centred mindset*" (Marcia Harris, Director of Strategic Initiatives, Office of Responsible AI at *Microsoft*). Stemming from their governance foundation, *Microsoft* has created two distinct tools we are highlighting here: Transparency Notes and Error Analysis.

2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it embedded/integrated in the AI product development processes (of the user of the tool)? How does it support the continuous monitoring of an AI product during usage?

AI principles are put into practice at *Microsoft* as part of a broader effort. *Microsoft*'s Responsible Resources support the responsible use of AI at every stage of innovation, including the assessment, development and deployment phases. Beyond the resources tailored to the different AI lifecycle stages, *Microsoft* has developed toolkits that help to integrate features into AI systems by theme. The Responsible AI Toolbox was released as an open-source framework to provide tools for the largest possible number of users. The toolbox includes elements from the area of error analysis, interpretability, fairness, counterfactual analysis, and causal decision-making, among others.

Table 6. The responsible AI principles

Fairness	**Reliability and safety**	**Privacy and security**
AI system should treat people fair	AI systems should perform reliably and safely	AI systems should be secure and respect privacy
Inclusiveness	Transparency	Accountability
AI systems should empower and engage people	AI systems should be understandable	People should be accountable for AI systems

Transparency Notes help customers and other stakeholders understand the functioning of an AI platform technology, to present the choices system owners can make to influence the system's performance and behaviour and to emphasise the importance of thinking about the technology, all affected stakeholders, and the context where the system is deployed. Transparency Notes do not concentrate on the model or dataset themselves but rather at a higher system level. In general, they intend to also "soften" the transition from development to deployment, highlighting capabilities and limitations and calling attention to mitigations that the customer can put in place when deploying the system. The tool has a clear, non-technical approach that is easily understandable for a larger audience (customers, implementers, regulators, media, etc).

As part of the Responsible AI Toolbox, the *Error Analysis tool* helps to analyse, understand, debug, and improve the predictions/decisions of a model based on the errors it may make. Importantly, it helps identify the "blind spots" where the model makes mistakes.

For example, the Error Analysis tool helps data scientists to potentially showcase fairness issues by assessing the level of error rate for subsets of a dataset compared to an overall benchmark error rate, and to visualize the distribution of errors.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

Transparency Notes are created for all AI platform systems at *Microsoft* and are ideally enhanced and used throughout the systems' development phase. *Microsoft* has put in place a dedicated team that provides "coaching" on how to develop these notes, provides user-friendly templates and guidance that are internally available at *Microsoft*. Although Transparency Notes can require significant effort create and maintain, there is a clear alignment in the on their value. There has also been an important learning process through the years that have resulted in improved templates and training that benefits all stakeholders involved in the process.

The Error Analysis tool can be used through the whole AI lifecycle to both identify and to help understand why and when errors occur (e.g., is the data properly representing all demographics) to continuously improve the models. Since it is a more technical tool compared to the Transparency Notes, it necessitates some level of data science knowledge and the understanding of the context it is used in. The tool is open source. However, is not released on its own, but it is part of the overall Azure platform and ecosystem.

4. Example case study

There are over 15 Transparency Notes published as of March 2022. One such example includes the Transparency Note for the Intelligent Recommendation service, which allows customers to build recommender systems using their business data and case study. For example, a customer can use the Intelligent Recommendation service to build a recommendation engine for their online store. In this system's Transparency Note, they highlight best practices (e.g., not using demographic data as an input for recommendations) and outline some limitations and corresponding mitigations (e.g., personalized recommendations for new shoppers may not be as robust due to a lack of historical interaction data. Instead, Intelligent Recommendations service can only generate recommendations for this shopper based on browsing of products from the current session). This Transparency Note can help customers who are considering using Intelligent Recommendation service understand if it will

meet their business needs, and it can also help developers as they integrate the platform solution into their end user experience.

Machine learning practitioners can use Error Analysis to gain a deeper understanding of model failure distribution and quickly identify erroneous cohorts of data. Often, error patterns may be complex and involve more than one or two features. Therefore, it may be difficult for developers to explore all possible combinations of features to discover hidden data pockets with critical failure. One of the demos available to try Error Analysis walks through a case study for evaluating a model to provide house sellers with advice on how best to price their houses in the market. Error Analysis has a tree visualization that partitions the benchmark data into interpretable subgroups, which have unexpectedly high or low error rates to gain insight as to which features could effect the error rate for underestimating the actual price, as an inaccurate price could impact financing and sale of the house. This is beneficial in providing an error identification view and understanding where the model production may fail.

5. What are suggestions to take into account while creating an AI quality assurance tool?

Based on the lessons learned from *Microsoft*'s Responsible AI governance framework as well as from the development process and use of the Transparency Notes and the Error Analysis tool, the following are key learnings:

- **Responsible AI policy and tools need to be well aligned with the organisation's governance structure and its general mission**; otherwise, it may be challenging to successfully translate them into practice;
- **Working with diverse teams is critical for tackling issues related to AI governance and responsible AI development and usage**, since the issues are very complex and are at the intersection of several different disciplines;
- **Allocating full-time personnel in the organisation to AI governance is key**. Otherwise, people may make suboptimal trade-offs on how they allocate their time in dealing with AI governance issues, which will often lead to inefficient outcomes;
- **Several tools are just focusing on one aspect of responsible AI**. This can create a "tools fatigue" and might lead to fragmentation and eventually lower impact. For this reason, a possible solution is, as also chosen by *Microsoft*, to have all the different tools under one dashboard to make sure modules are easily and freely available to be able to achieve a higher overall impact;
- **Users wish and need to know at what point a model can be considered acceptable from a "responsibility perspective"**, although the answer to this question can be very specific for each case study. Therefore, stakeholders need to be further trained to be able to translate business (and societal) needs into responsible technical requirements;
- **Very few responsible AI tools concentrate on risk mitigation**. Most existing tools focus on diagnosis and identification. Mitigation is a complex issue and can be investigated in several different ways;
- **A decision needs to be taken between providing open-source or licenced tools.** There is a trade-off between the two approaches with advantages and some possible inconveniences that all need to be jointly considered.

AI Quality Assurance tool at NEC

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

NEC Corporation (*NEC*) is a Japanese multinational, headquartered in Tokyo. The company provides IT and network solutions, including cloud computing, AI, IoT platforms, communication equipment and software to business enterprises, communications service providers and government agencies. To support the company's overall trustworthy AI framework, *NEC* ensures the integration of reliable AI and Human Rights Principles at all phases of its AI development process. In addition to facilitating compliance with relevant laws and regulations around the globe, *NEC* is enacting the principles to prevent and address human rights issues arising from AI utilization.

AI and Human Rights Principles are overviewed by in-house governance units, including the Digital Trust Business Strategy Division (DTSD) responsible for supervising the implementation of the trustworthy AI framework into day-to-day practice based on the principles. The Division's main tasks relate to institutional coordination and employee education while it is in direct contact with AI-related business units and conducts a dialogue with external experts and organisations to integrate stakeholder feedback into practice. Members of the governance unit have mainly technical, business, ethical and legal profiles. *NEC* executives are also involved in the unit's activities to have a proper understanding of human rights issues, especially in the case of AI.

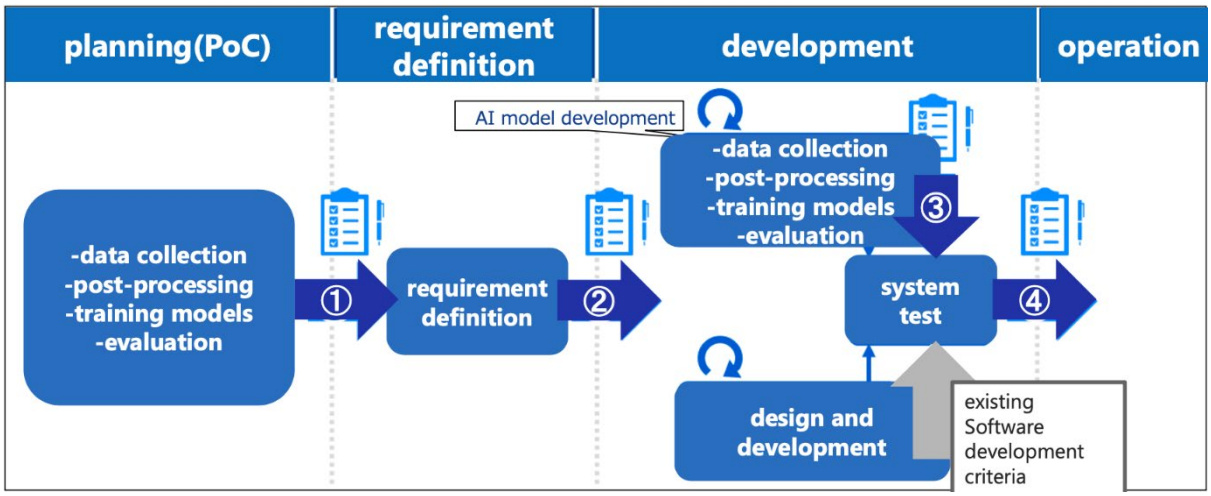
To support its trustworthy AI framework, *NEC* has developed a quality assurance tool[35] based on guidelines provided by the Japanese government as well as a Japanese private sector consortium. *NEC*'s AI quality assurance checklist has been developed as an extension of the quality assurance checklists for software development that previously existed. Guidelines for Quality Assurance for Machine Learning-based AI. The intention for developing the AI quality assurance tool was to create a practical instrument to support engineers at *NEC* with specific criteria on quality assurance along the AI product development process. The tool supports robustness, safety and accountability principles through providing guidance for instance on how to handle data, build models or reach an agreement with customers on specific issues.

2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it embedded or integrated in the AI product development processes (of the user of the tool)? How does it support the continuous monitoring of an AI product during usage?

NEC has structured the development of its AI projects into four different phases along the AI lifecycle, corresponding to planning, requirements definition, system development, and operation phases (see Figure 4). In the planning phase, a proof of concept is agreed upon and brought to the requirements definition phase, where the system is specified in detail. The system development phase involves AI model development and testing, leading to the final operational phase.

A project is defined at *NEC* as a business case to develop an AI solution for a particular client, with every project going through the internal AI quality assurance tool. The tool includes checkpoints at the end of each phase of the AI product's lifecycle. At each checkpoint, a set of criteria are examined and reflected upon in the form of a checklist. There would be for instance questions asked about the availability of sufficient samples to train AI models, the quality

Figure 4. Development of AI products along the AI lifecycle at NEC



of the data, or about the definition of errors and imperfections in customer-owned datasets as well as the negotiation process with the customer for removing inappropriate data. All together, the tool includes more than one hundred questions, and around 20 to 60 of them are evaluated at each checkpoint.

There are no "one type fits all" checklists used at *NEC*. Instead, a customised version is being created by project managers for each project. However, there are also some core questions that raise key issues and are in general important for all case studies. Lack of satisfying some of the questions will not mean that the project has failed (on average 80% of the projects passed 67% of the checklist questions). Instead, a continuous dialogue between the *NEC* team and the client allows for adjustments in a mutually acceptable manner. Final judgement calls about whether to "overwrite the guidelines" can be made by the product manager, and in such a case *NEC* engineers make sure clients clearly understand and accept the related risks.

As part of the continuous improvement of the checklist tool, both external and internal AI engineers and data analysis team members play an important role by providing feedback on the quality assurance tool. Feedback is being given in a structured way in the form of recommendations at the end of each project. Such recommendations would concern for instance how the checklist could be customised to each case or would reflect on the specific needs clients have formulated and the way these should be reflected in the tool. Besides such customization-related improvements, *NEC* is currently strengthening the ethical aspects in the checklist by promoting a discussion among relevant business units including DTSD.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

The main users of the guidelines are product managers through the four phases of the AI development process, but different experts also interact with the tool during the AI lifecycle. AI consultants are active more in the planning and ideation phases, data analytics experts support the requirements definition and validation phases, while AI architects take part in the system development and operation phases.

In addition to having tools such as the AI quality assurance one, it is also important that users of the tools as well as people involved in AI have strong related skills. To support the AI skill development of its employees, *NEC* has also

Table 6. Development of AI products along the AI lifecycle at NEC

AI development phase	Experts involved at each phase
Planning	Analysis consultant, AI analytics coordinator
Requirements definition	Data analytics expert, AI analytics coordinator, system engineers
System development	AI architects, system engineers
Operation	Operation and maintenance staff leaders

elaborated a Literacy Programme tailored to educational backgrounds, experience and roles in the organisation. Different types and levels of data literacy and project-management skills are required on behalf of internal users of the tool at *NEC* depending on their role and responsibilities. Guidance is also taken from the Japanese Data Scientist Society which outlines the type of skills people involved in data and AI are expected to have – largely grouped around general business, data science and data engineering skills.

Finally, this specific quality assurance tool is customized to *NEC*'s business processes and thus is only for *NEC*'s internal use. However, several other companies have also developed their own AI development guidelines.

4. Example tool improvement based on a specific use case

An example of how feedback from a specific use case was used to improve the quality assurance checklist comes from an *NEC* project where the company developed and provided an image recognition AI solution to its business client that automatically detects and alerts about defective products in the customer's manufacturing line using the products' images. The feedback from this use case resulted in a revision of one question in the quality assurance checklist. There used to be a question in *NEC*'s quality assurance tool that said "check if there is no label data leakage (OK/NG)". In the training process, data for extracting features must not contain information on the answer (e.g., if that output is true or false in the case of a classification task). And if this condition does not hold (i.e. if the data contains information on the output), we have the issue of label data leakage. This question in the checklist was first developed without considering the application of AI in image recognition. When the training data consisted of numerical and text data, the label data leakage was explicitly detectable and was easy to remove from the training datasets.

However, label data leakage was found to occur for image recognition when a true/false label was included into an image itself in the form of a true/false character image. Thus, recognition of such a situation called for some additional check processes:

> (1) human inspection of every image to detect label data leakage and,
>
> (2) additional negotiation with the customer for checking the quality of training images.

As a result, the revised question in the checklist now has an added phrase "check and confirm with customer members not to contain label information even if it is image data, because it sometimes contains OK/NG label information in the image data itself".

5. What are suggestions to take into account while creating an AI quality assurance tool?

NEC has acquired considerable experience in implementing the AI quality assurance tool. As a result, the following points are suggestions to be taken into account during the development process of similar tools:

- **Building on existing tools (e.g., in this case software quality assurance ones) and knowledge during the development** of an AI (quality assurance) tool can be beneficial, especially to build trust and acceptance of the tool by the organisation (for example by engineers who would use the tool);
- **Finding the right balance between pre-definition of the tool and customisation** is important. Some appropriate customization of the tool to each project increases efficiency and effectiveness since AI system requirements can be different for each use case. However, some fundamental requirements can be identical for all AI development projects;
- **Human oversight and judgement on "overwriting" the guidelines of the tool** provides flexibility and helps customisation. This flexibility is provided based on case-by-case decisions of the product managers to eventually "overwrite the guidelines" if needed to adapt to the special circumstances of each project while informing the clients about possible related risks. A process to decide and communicate any diversions from checklists needs to be in place;
- **Continuous improvement** of the tool is key as AI technology and related use cases are constantly evolving and this needs to be reflected by the tools used along the development process of AI products. Feedback from stakeholders, such as clients from outside the organisation and from internal users of the tool is an effective way for continuous improvement. Hence feedback and continuous improvement processes need to be in place;
- **Some level of related skills and data literacy is needed** on behalf of all stakeholders involved in using the tool. This needs to be ensured through training tailored to the educational background, experience and role in the organisation;
- **Tools and best practices can be shared across organizations**. For example, *NEC*'s contribution to the development of skills and data literacy is shared with their customers as well as with broader community members. For instance, a set of courses called "*NEC* Academy for AI" is offered as part of developing AI literacy programmes. Several types of skills and expertise corresponding to every phase of the AI lifecycle are classified, and respective training programs that are standardized and customized to each phase of expertise are developed. *NEC*'s work has also contributed to the Japanese government's initiative on AI education "Japan Inter-University Consortium for Mathematics & Data Science Education".

Responsible AI Toolkit and the Bias and Fairness tool at PwC

1. What is the AI tool or practice in question? What does the tool aim to achieve? Why did you decide to develop a tool in this area – what are the motivating forces? What are the business implications?

PricewaterhouseCoopers (*PwC*) is a multinational professional services network of firms operating under the same brand. It is one of the Big Four accounting firms, being present in more than 150 countries with around 300.000 people.

Starting from 2016, the need for more context and perspective on how to deal with AI governance, models and related risks was increasingly articulated by *PwC* clients. The first requests came from the financial and tech industries while demands from other sectors soon followed. By this time, there was already a significant amount of research published on ethical AI while the European Union and international organisations (e.g., OECD, IEEE) were developing their approach and principles about this new space. However, there was still an important gap between academic research, principles and what the companies needed in practice: *"Our clients were asking questions about the level and nature of risk business executives and data scientists face and the possible actions to mitigate these risks" (Anand Rao, PwC Global AI Lead).*

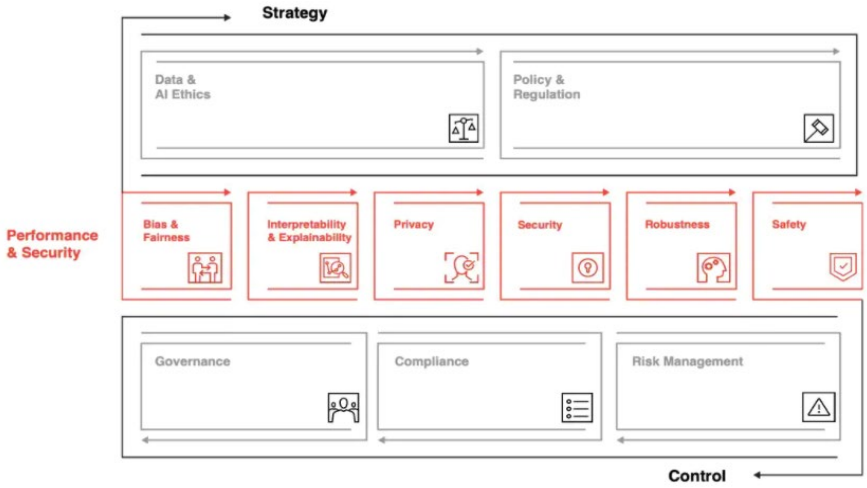
Therefore, the Responsible AI Toolkit aimed to bridge the gap between principles and practice, while ensuring not to develop a set of tools that are deployed without considering the broader socio-technical context. The Responsible AI Toolkit is, therefore, more than a technical solution, as it also considers the interaction between people and AI systems.

2. How does the tool work? At what point in the AI lifecycle should the tool be used? How is it embedded/integrated in the AI product development processes (of the user of the tool)? How does it support the continuous monitoring of an AI product during usage?

The Responsible AI Toolkit (see Figure 5) is a set of customizable frameworks, tools and processes designed to help harness the power of AI ethically and responsibly – from strategy through execution. It includes three different modules centred around:

1) Strategy;
2) Performance and Security;
3) Risk management, Compliance and Governance.

Figure 5. The structure of the Responsible AI Toolkit[36]



The *Strategic module* identifies and compares the different policy and regulatory aspects concerning AI principles (e.g., published by the EU, international organisations, professional organisations, and companies). *PwC*'s intelligent tool, as part of the strategic module, helps clients to assess their approach on specific aspects (e.g., transparency) within the existing policy and regulatory landscape. The tool also looks beyond AI principles and covers data and privacy policies since these are important components of AI systems.

The *Performance and security module* includes technical tools that relate to six "core" trustworthy AI principles that are the most often referred to by different strategic documents. These six technical tools (see Figure 5) can help assess the successful implementation of these AI principles in practice.

The *Risk management, compliance, governance module* is based on key lessons and practices from the financial sector, such as the risk tiering approach to mitigate risks used in that sector. The module defines the duties of the "three lines of defense":

1) Data scientists;
2) Compliance;
3) Internal audit concerning the use case of the model.

The toolkit also includes general tools related to readiness and risk assessment, which provide evaluation and guidance to define the appropriate level of different risks and to associate the right actions for mitigating these risks. The toolkit also includes frameworks and processes defining the governance, decision processes and deliverables at all stages of the AI lifecycle, including the initial business decisions, model development, deployment, retraining, continuous learning, and monitoring.

3. Who is involved in using the tool during the development of AI products? What are the roles needed and what are their responsibilities? Does the tool require specific skills or training to use? If so, how is this training done? Does the tool need the support of specialists to implement and use, or once developed it can be used by all? Is the tool business-facing or customer-facing? Is it usable by other companies or sectors?

Based on *PwC*'s experience the tool is being used in three different contexts. On the policy level, mainly by privacy officers and legal officers looking for guidance to incorporate responsible AI governance into strategic level documents and practices of the organisation. At the engineering and AI system development or deployment level, data scientists use the tools and practices to incorporate requirements of responsible AI procedures into their work. Finally, risk management and compliance divisions use the tools to manage and reduce the level of all type of risks including the new types of risks related to the field of using data-intensive technologies such as AI.

The Responsible AI Toolkit is not open source. *PwC* made this choice to ensure the proper usage of the toolkit, which requires users to also receive guidance on how to use it in alignment with their context and use cases. This reduces the risk of clients just using the tools for "ticking the boxes" and possibly misinterpreting or misusing it – e.g., to validate their models without following a strong process, or by not making appropriate trade-offs.

The Toolkit is modular, with new modules being constantly added for example for additional use cases such as natural language processing, image recognition, etc. Diversity of the team developing the Toolkit was an important consideration at *PwC*, to ensure inclusive product design and consideration of the broadest number of potential issues. The Responsible AI Toolkit team included around 40 people and 17 nationalities from many offices inside *PwC*.

4. Example use case

The *Bias and Fairness analyser tool* is part of the performance and security module of the Toolkit.

This specific tool highlights various aspects of AI fairness to consider and analyse when developing an AI system. It is built also based on *PwC*'s experience with the fair lending model valuation for financial services, with AI principles being incorporated today into a broader class of models than were traditionally considered before. The analyser can test any model against 30+ definitions of fairness and can provide a score showing how much the model is fulfilling each one of them. It is then up to the team developing the AI system to define the type of fairness it intends to satisfy.

The tool also helps identify different trade-offs. For instance, while allocating loans, a certain group of people might be discriminated against based on their age and likelihood of defaulting on their payments. In this case, if fairness is improved it is probably also impacting business costs and revenues. The Bias and Fairness analyser helps assess such trade-offs and gives further elements to business users for them to make decisions about the final system.

5. What are suggestions to take into account while creating an AI quality assurance tool?

Based on the lessons learned from the *PwC* Responsible AI Toolkit and the specific Bias and Fairness module, the following suggestions can be highlighted when developing or using this, or other similar toolkits:

- **AI principles can have several different interpretations and implementations**, and it may not be possible to satisfy all of them at the same time. Alternatives for making the right decisions need to be provided to users, with the final choice made following a strong process and ideally by involving a diverse group of stakeholders;
- **It can be misleading to test AI solutions only through using technical, quantitative tools**. The specific socio-technical context needs also to be taken into consideration when evaluating responsible models;
- **AI principles need to be operationalised** to fill the gap between principles and practice. This is not only the duty of the technical professionals (e.g., data scientists). Instead, "end-to-end" AI governance needs to be implemented along the whole AI lifecycle and across all levels of an organization,
- **There needs to be clarity about reporting and escalating AI risks and important decisions to the C-level.** Business leaders need to receive the appropriate level of information and knowledge related to AI systems, to fully understand risks and make the right decisions;
- **Proper AI governance needs to be embedded into the organisation**. This needs to be ensured at all levels of the organisation;
- **Data and AI models need to be evaluated jointly** since data is a fundamental part of an AI system. The evaluation of responsible AI systems must adopt data related best practices.

Footnotes

1. Companies which participated to the study include *AXA*, *Amazon Web Services* (*AWS*), *IBM*, *Facebook*/*Meta*, *Microsoft*, *NEC* and *PwC*.
2. We would like to thank the experts from participating companies who contributed to the development of the case studies.
3. Theodoros Evgeniou is Professor at INSEAD, World Economic Forum Partner on AI, member of the OECD Network of Experts on AI, BCG Henderson Institute Advisor, and Co-Founder and Chief Innovation Officer of Tremau, a B2B SaaS company whose mission is to build a digital world that is safe beneficial for all. He has been working on AI for more than 25 years. He holds four degrees from MIT.
4. Pal Boza is a Senior Researcher at INSEAD, and Co-Founder and Chief Operating Officer at Tremau. Pal holds and Executive MBA from INSEAD and master's degree from ENA.
5. "Evaluating a Methodology for Increasing AI Transparency: A Case Study" by Piorkowski et al, 2022, https://arxiv.org/abs/2201.13224
6. https://www.axa.com/en/insights/the-fairness-compass-a-groundbreaking-step-forward-for-trustworthy-ai
7. https://github.com/axa-rev-research/fairness-compass
8. https://www.axa.com/en/insights/the-fairness-compass-a-groundbreaking-step-forward-for-trustworthy-ai
9. See for example "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI", Sandra Wachter, Brent Mittelstadt, Chris Russell, Computer Law & Security Review, 2021
10. https://www.wired.com/story/social-media-ceo-hearing-cant-defend-business-model/
11. ICO and Alan Turing Institute. Explaining decisions made with AI. 2020, https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/(accessed on the 20th of August 2020)
12. https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/ (accessed on the 20th of September 2020)
13. https://about.fb.com/news/2021/08/widely-viewed-content-report/ (accessed on the 20th of September 2020)
14. https://about.fb.com/news/ (accessed on the 20th of September 2020)
15. https://about.fb.com/news/2019/07/addressing-sensational-health-claims (accessed on the 20th of September 2020)
16. https://about.fb.com/news/2018/05/inside-feed-news-feed-ranking (accessed on the 20th of September 2020)
17. https://transparency.fb.com (accessed on the 20th of September 2020)
18. https://transparency.fb.com/policies/community-standards (accessed on the 20th of September 2020)
19. https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021 (accessed on the 20th of September 2020)
20. For users who prefer not to have the algorithm decide which content they see, we also have the Most Recent feed view - which shows chronological content as it comes into the user's feed. Currently, users have to manually select it, but it is another option. https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/

21. Meta (https://tech.fb.com/news–feed–ranking/)
22. https://about.fb.com/news/2021/09/content-distribution-guidelines/
23. https://about.fb.com/news/2019/03/why-am-i-seeing-this/
24. https://www.crowdtangle.com/ (accessed on the 20th of September 2020)
25. For additional information see: https://aif360.mybluemix.net/; https://aix360.mybluemix.net/; https://aip360.mybluemix.net/; https://art360.mybluemix.net/; https://uq360.mybluemix.net/
26. Source and further information: https://aifs360.mybluemix.net/governance
27. Richards et al., A Methodology for Creating AI FactSheets, 2020, https://arxiv.org/abs/2006.13796
28. https://aifs360.mybluemix.net/governance
29. Richards et al., A Methodology for Creating AI FactSheets, 2020, https://arxiv.org/abs/2006.13796
30. Further information about example FactSheets developed by IBM available here: https://aifs360.mybluemix.net/examples, accessed on 02.06.2021
31. Detailed version of the Audio classifier example FactSheet availible here: https://aifs360.mybluemix.net/examples/max_audio_classifier, accessed on 02.06.2021
32. Detailed version of the Mortgageevaluator governanve example FactSheet availible here: https://aifs360.mybluemix.net/examples/hmda, accessed on 02.06.2021
33. Source and further details: Hind et al., Experiences with Improving the Transparency of AI Models and Services, 2019, https://arxiv.org/abs/1911.08293v1
34. https://aifs360.mybluemix.net/methodology
35. Guidelines for Quality Assurance for Machine Learning-based AI
36. PwC

Business at OECD (BIAC)
13-15, Chaussée De La Muette – 75016 Paris
Tel: + 33 (0) 1 42 30 09 60
@BusinessAtOECD | businessatoecd.org