EUROPE | JANUARY 2021

AI Impact Assessment: A Policy Prototyping Experiment

NORBERTO NUNO GOMES DE ANDRADE

VERENA KONTSCHIEDER

About Open Loop

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

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

This report presents the findings and recommendations of the Open Loop's policy prototyping program on AI Impact assessment, which was rolled out in Europe from September to November 2020.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Cite this report

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

Acknowledge-ments

This policy prototyping program was co-designed and facilitated by Facebook and the consulting firm Considerati. A special thank you to the Considerati team, in particular to Dr. Bart Schermer and Joas van Ham, for their invaluable contribution to this project.

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

Allegro Israel
Evo Italy/Great Britain
Feedzai Portugal
Irida Labs Greece
Keepler Spain
NAIX Technology Germany
Reface AI Ukraine
RiAtlas Italy
RogerVoice France
Unbabel Portugal

We would also like to thank the many experts that participated in the AI Impact assessment policy prototyping program workshops, namely Eva Maydell, Sofia Ranchordas, Emilia Gómez Gutierrez, Jochen Mistiaen, Martin Ulbrich, Bojana Bellamy, Giuseppe Fenza, Pedro Bizarro, Ramin Karbalaie, Jason Li, Moses Guttmann, Ariel Biller, Ramiro Manso, Ezequiel Paura, Roman Mogylnyi, Oles Petriv, Alon Lavie, Olivier Cuzacq, Pedro Saleiro, Igor Carvalho, Michal Schwartz, Luca Romanelli, Evangelina De Luca, Thomas Charisis, Michel van Leeuwen, Duuk Baten, Roffel Sweitze, Claudine Vliegen, Nathalie Laneret, Edo Haveman, Janne Elvelid, Nicolas de Bouville, Lawrence Muskitta and all other attendees.

Contents

Contents

Executive Summary

This report presents the outcomes of the Open Loop policy prototyping program on Automated Decision Impact Assessment (ADIA) in Europe. Open Loop is a collaborative initiative supported by Facebook to contribute practical insights into policy debates by prototyping and testing approaches to regulation before they are enacted.

Facebook partnered with 10 European AI companies to co-create an ADIA framework (policy prototype) that those companies could test by applying it to their own AI applications. The policy prototype was structured into two parts: the prototype law (drafted as legal text) and the prototype guidance (drafted as a playbook). The goal was to derive evidence-based recommendations relevant to ongoing policy debates around the future of AI regulation.

Participating companies were asked to select an Artificial Intelligence (AI)/Machine Learning (ML) application that would produce effects or have an impact on people and simulate the application of the ADIA framework on that particular application. Participating startups were asked to provide their initial feedback on the prototype law, then simulate the implementation of the ADIA process based solely on its contents. Participants later received a playbook providing them with a step-by-step methodology, along with a list of values relevant to AI/ML and automated decision-making (ADM) and a taxonomy of harms and examples of mitigating measures, and were asked to provide feedback on how this additional guidance would have changed their implementation. Throughout the program, participants shared their experiences through a mobile ethnography application and dedicated workshops.

The results of this initial policy prototyping program clearly demonstrated the value of implementing an ADIA framework as a tool for identifying and mitigating risks from AI/ADM systems. The results also highlighted the need for clearly defined guidance on how to implement that framework practically and the importance of ensuring consistency with existing obligations like GDPR's Data Protection Impact Assessment (DPIA) requirements.

This program further demonstrated that a procedural approach to risk assessment, where organisations identify, assess, and mitigate risks by following a series of steps, indicative criteria, and examples, can be an adaptable alternative to a prescriptive regulatory approach applied to specific business sectors or intended uses. A step-by-step risk assessment approach, complemented by a set of examples of risks and taxonomy of values, proved to help organisations assess risks based on the specific context and impact of their proposed AI uses while taking into account the dynamic and iterative character of AI.

The ADIA framework

The prototype ADIA framework that we tested aims to make AI developers and users (organisations deploying ADM systems) aware of the risks their applications may pose and enable them to find ways of mitigating these potential risks. To achieve this goal, the framework requires actors to perform risks assessments for their AI/ADM application. The ADIA process outlines four requirements to be met by the organisations (users) deploying the AI/ADM system:

- **Users are able to identify what risks their applications may entail for the rights and freedoms of subjects;**
- **Users are able to determine how significant these risks are (e.g. high or low);**
- **Users are able to formulate mitigating measures to these risks;**
- **Users are able to adequately assess whether these measures remove the risks or reduce them to an acceptable level (residual risk).**

Outcomes
Based on the ADIA simulation and the feedback of participants on the ADIA framework, we evaluated the legal text against three criteria: policy understanding, policy effectiveness, and policy costs.

1 **Policy understanding**

2 **Policy effectiveness**

3 **Policy costs**

Regarding policy understanding, we concluded that the prototype law was sufficiently clear in its wording for users to develop a basic understanding of what was required of them, although they still had significant open questions about how exactly to comply, and only half were confident that they could do so based solely on the guidance of the legal text. This points to the need for clear supplemental guidance, beyond legal text, detailing specific instructions and expectations. Some participants were also unclear on how they would be categorized under the prototype law's definitions of relevant actors, highlighting both the complexity of the AI landscape and the need for greater clarity in how the law parses it.

Regarding policy effectiveness, we concluded that the prototype law was helpful overall in prompting participants to fulfill the intended requirements of identifying risks and formulating mitigations to address them. However, there was a wide variance between companies in the types of risks they considered, with most focusing solely on risks related to the design and operation of their system such as dataset bias and performance issues – i.e. functional risks – as opposed to a broader set of risks related to the ethical application of ADM systems, and the societal effects of these decisions such as impact on human well being, fairness, human interaction, end user autonomy, or overreliance on AI/ADM systems – i.e. structural risks.

This highlights the need for policymakers to be very clear about what types of risk they are and are not attempting to address in any risk assessment requirement, and also highlights the challenges of expecting companies to broadly identify and mitigate every conceivable kind of risk.

There was also a gap in terms of participants completing the second and fourth steps contemplated by the process- gauging risk severity to inform mitigation decisions, and assessing residual risk after mitigations – indicating a need for greater clarity on that point in the prototype law and guidance.

Like many regulatory requirements, there are policy costs involved in complying with the requirements of the framework. While the investment of time and resources to implement the framework was significant, there was no indication in this limited test that performing an ADIA would overburden the participants. This was especially true for participants already complying with the GDPR's DPIA requirements, where there is some overlap with the prototype law's requirements. Ideally, there would be a proper integration between ADIA and DPIA requirements in law to avoid duplicative costs for developers and users.

The introduction of the playbook provided the participants with additional guidance and examples of potential risks and values. According to the feedback, the playbook helped participants translate the prototype law to their own contexts and made implementation more straightforward. Feedback by participants showed that all of them would actually change their risk assessment after reading the playbook. This further demonstrates the need for additional guidance through operationalization and examples for a shared, practicable understanding of an ADIA requirement.

Recommendations

Based on the results of the prototyping exercise and the feedback on the prototype law and playbook, we would advise lawmakers formulating requirements for AI risk assessments to take the following recommendations into account:

- **Focus on procedure instead of prescription as a way to determine high-risk AI applications.** The findings show the importance of codifying a risk assessment procedure. This procedural approach – unconstrained by prior sectoral determinations and complemented by a set of examples of risks and taxonomy of values – will do a better job at helping organisations assess risks based on the specific context and impact of proposed AI uses. The higher level of uncertainty and complexity of the types of risk posed by AI/ADM systems requires robust step-by-step procedural approaches to risk assessment, complemented with operational guidance, rather than an approach anchored on rigid classifications based on the sector in which AI is being utilized.

- **Leverage a procedural risk assessment approach to determine what is the right set of regulatory requirements that apply to organisations deploying AI applications.** Rather than applying an entire set of regulatory requirements by default and regardless of the type of AI application, its context, and actual risks, the procedural approach allows for a more balanced and appropriate application of regulatory requirements in response to identified risks: human oversight, explainability, rights of redress, monitoring, and disclosure requirements, amongst others. Through such an approach, statutory requirements are assigned not in bulk, but in accordance with the specific AI application in question and the level and extent of the risks assessed, alongside the calculus of the benefits that application brings.

- **Provide specific and detailed guidance on how to implement an ADIA process, and release it alongside the law.** The positive impact of the playbook's additional guidance on the participants' risk assessments shows the need for similar guidance accompanying legal requirements. Guidance such as that provided through the ADIA playbook helps overcome the inherent ambiguity of norm-based regulation (which is needed for the policy to be technologically neutral), and through taxonomies and examples helps identify previously unknown aspects of an ADM system. The demand for additional guidance also confirms the need for a tighter calibration and coordination between different governance instruments: hard law, soft law, and co-regulation.

- **Be as specific as possible in the definition of risks within regulatory scope.** The results show that risks related to the functioning of AI systems (how they are built and operate) are easier to identify than risks related to the application of those systems and their broader consequences for individuals and society. In particular, the participants' feedback on the guidance on values and harms, provided through the playbook, showed that it was difficult for them to understand how their products or services may implicate abstract values (for instance human autonomy). To avoid such uncertainty, we urge policy and lawmakers to work with academia, civil society, and industry to clearly specify the types of risks and harms that are to be identified in a systematic manner and mitigated in an effective way. Providing playbooks like the one for our ADIA framework, defining specific values to weigh and providing a clear taxonomy of harms to consider, can be a good first step to reduce this uncertainty and avoid the burden of companies trying to identify and solve every possible moral implication of their AI-based products and services.

- **Improve documentation of risk assessment and decision-making processes by including justifications for mitigation choices.** Deciding and documenting how to mitigate risks posed by AI systems needs to be part of any AI risk assessment process, and is a fundamental element informing the overall AI risk-based approach. Based on the feedback received by our program participants, it would be helpful if users (deployers) of an ADM system also described in their ADIA why particular risk-reducing measures were taken (and others not), and how these measures reduced the risk to an acceptable level (or removed it altogether). The reasons for accepting any residual risk should also be included in the ADIA. Providing these further insights

on the value and effectiveness of the risk mitigating measures selected would help determine the right set of regulatory requirements applicable to the AI application in question, and bring greater clarity as to how tensions amongst values affected by AI/ADM are resolved.

- **Develop a sound taxonomy of the different AI actors involved in risk assessment.** When regulating AI/ADM, lawmakers must be cognizant of the complex landscape of actors developing, deploying, using, and being impacted by AI/ADM. The development of such taxonomy is important for two main reasons: to appropriately assign the tasks of identifying, assessing or mitigating risks; and to better understand the group of stakeholders being affected by AI/ADM.

- **Specify, as much as possible, the set of values that may be impacted by AI/ADM and provide guidance on how they may be in tension with one another.** When implementing a requirement to do a risk assessment for AI/ADM, it is important to clarify to the entities called to perform the ADIA what is required of them. In particular, guidance and explanation on values that may be affected by AI/ADM and value tensions that may arise are very helpful. Greater clarity around the values that should be weighed when balancing the risks of a particular technology or mitigation approach against its benefits would in turn enable better decision-making, and better documentation of why particular decisions were made.

- **Don't reinvent the wheel; combine new processes with established ones, improving the overall approach.** In many cases, there is an overlap between the ADIA and the GDPR DPIA requirements. In order to avoid duplicative work and costs, a proper integration between ADIA and DPIA requirements in law is necessary.

This project was not only meant to test the idea of ADIAs, but also the idea of policy prototyping itself. Based on the helpful results of this proof of concept, we intend to continue with similar projects. In addition to the AI policy-specific recommendations above and based on our positive experience, we would further urge policy-makers to support or participate in similar projects to test novel approaches to regulating complex technology policy issues before codifying those approaches in law, whether in partnership with the Open Loop initiative or otherwise.

Introduction

A risk-based approach to AI: An emerging regulatory trend

01

Introduction

A risk-based approach to AI: An emerging regulatory trend

The need to identify and assess risks posed by Artificial Intelligence (AI) and Machine Learning (ML) systems has emerged as one of the mainstream approaches to AI governance and regulation. According to this view, regulatory requirements should only be applied to AI or automated decision-making (ADM) systems and applications that present a certain level of risk.[1] Governments, international institutions, standard organisations, businesses, academics and civil society institutions have either based their approach to AI governance on "risk" or supported such an approach.

- The European Commission has followed a risk-based approach in the AI regulatory framework outlined in their "White Paper on Artificial Intelligence – A European Approach to excellence and trust", stating that "[a] risk-based approach is important to help ensure that the regulatory intervention is proportionate."[2]
- The OECD, in its Recommendation on AI, asserts that "AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems."[3]
- UNESCO also emphasizes a risk-based approach for AI in its evolving "Recommendation on the Ethics of Artificial Intelligence", encouraging Member States "to introduce impact assessments to identify and assess benefits, concerns and risks of AI systems,as well as risk prevention, mitigation and monitoring measures."[4]
- The IEEE, in the same vein, underlines the importance of a systematic risk analysis and management approach in their Ethically Aligned Design recommendation for autonomous and intelligent systems.[5]
- On the governmental side, 14 EU countries signed a non-paper entitled "Innovative and Trustworthy AI: Two sides of the same coin" supporting a risk-based approach towards AI, and stating that "[w]here specific situations related to risks to individuals or society stemming from the use of AI are not tackled by existing legislation, we need to address these by a risk-based legislative framework protecting existing public values and fundamental rights."[6]
- In the US, the Guidance for Regulation of Artificial Intelligence Applications issued by the Office of Management and Budget (OMB) states that "[r]egulatory and non-regulatory approaches to AI should be based on a consistent application of risk assessment and risk management across various agencies and various technologies."[7] And the proposed Algorithmic Accountability Act would require entities to perform automated decision system impact assessments of high-risk automated decisions.
- In Asia, Singapore's AI Governance Model Framework also follows a risk-based approach to AI governance, proposing a matrix to help organisations determine the level of human involvement in AI decision-making. That matrix lays out a number of factors that could be used as operational guidance to assess risk posed by AI systems: nature, probability, severity, and reversibility of harm, amongst others.

1. We have used the term automated decision-making and AI interchangeably throughout this policy prototyping program, and will do the same for this report for ease of reading.
2. European Commission (EC) 2020a (p.17).
3. Organisation for Economic Co-operation and Development (OECD) 2019 (Art. 1.4 (c)).
4. UNESCO 2020 (p. 13).
5. IEEE 2019.
6. Position paper on behalf of Denmark, Belgium, the Czech Republic, Finland, France, Estonia, Ireland, Latvia, Luxembourg, the Netherlands, Poland, Portugal, Spain and Sweden on innovative and trustworthy AI.
7. Office of Management and Budget (OMB) 2020

- In Canada, the Directive on Automated Decision-Making requires relevant Canadian federal agencies to conduct an Algorithmic Impact Assessment for any automated decision system (ADS) developed or procured to the extent that the ADS will be used to recommend or make an administrative decision about a client. The Canadian government, through a public-private partnership, subsequently has begun developing a model Algorithmic Impact Assessment tool that the relevant agencies could refer to (or use) in complying with the Directive on Automated Decision Making.

The risk-based AI approaches put forward also vary in terms of their connections to and interplay with existing legal frameworks. We have seen proposals that defend the incorporation of human rights concepts, frameworks and processes as the basis for AI risk assessments,[VII] along with proposals pointing to GDPR's Data Protection Impact Assessments (DPIAs) as models from which to develop and implement Automated Decision Impact Assessments (ADIAs). [VIII]

Within this emerging trend towards a risk-based approach to AI governance, there is a variety of perspectives on how to define and assess the risks posed by AI systems:

- Some risk-based approaches follow a binary "high-risk / low-risk" determination,[I] while others propose a multi-tier risk classification.[II]
- Some follow a prescriptive approach based on comprehensive and exhaustive lists of factors as criteria to identify risks (such as sectors, intended use cases),[III] while others recommend a procedural approach-based on a set of steps and/or questions meant to arrive at risk determinations through qualitative analysis, dialogue and reflection.[IV]
- Some AI risk approaches may rely on a quantitative type of assessment based on the calculation of risk scores,[V] while others advocate for a qualitative assessment based on collection of stakeholder inputs on risk.[VI]

Table 1: Risk-based approaches to AI regulation

Risk-based approaches to AI Regulation	Examples
Binary "high-risk / low-risk"	European Commission's White Paper on AI
Multi-tier risk classification	German Data Ethics Commission's 2019 opinion on algorithmic and data governance,
Prescriptive	European Commission's White Paper on AI
Procedural	AI HLEG Assessment List for Trustworthy AI (ALTAI) Singapore's AI Model Governance Framework, and its Companion Guide Considerati / ECP's Artificial Intelligence Impact Assessment IEEE's Standard 7010-2020: Assessing AI Impact on Human Well-Being Human Impact Assessment for Technology (Rafael Calvo et al) Open Loop Automated Decision Impact (ADIA) Framework
Quantitative	Canada's Algorithmic Impact Assessment tool
Qualitative	[see examples listed for procedural approach]
Human Rights driven	European Union Agency for Fundamental Rights' (FRA) Getting the Future Right: Artificial Intelligence and Fundamental Rights Center for Democracy and Technology's Response to EC White Paper on AI Data and Society's Governing Artificial Intelligence: Upholding Human Rights and Dignity AI and Big Data: A blueprint for a human rights, social and ethical impact assessment (Mantelero)
GDPR DPIAs aligned	Chair Legal and Regulatory Implications of AI of Université Grenoble Alpes Submission to the EC's White Paper on AI Private Accountability in the age of Artificial Intelligence (Katyal) Facebook's response to EC White Paper on AI

Despite the overwhelming consensus and ample agreement on the need and importance of a risk-based approach to AI, and despite the various risk assessment modalities put forward, the development of concrete and operational AI risk assessment frameworks is – with a couple of noteworthy exceptions – still lacking. For all of the writing about AI principles and governance, only a few concrete and operationalizable AI risk assessment frameworks have been proposed (see table 2), and none of them have garnered widespread consensus at this point.

Table 2: Operational AI risk assessment frameworks proposed to date

EU High Level Expert Group Assessment List for Trustworthy AI (ALTAI)
Singapore's Model AI Governance Framework and its companion guide, the Implementation and Self-Assessment Guide for organisations (ISAGO)
The Government of Canada's Algorithmic Impact Assessment (AIA)
IEEE 7010-2020 Assessment of AI Impact on Human Well Being
Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (AI Now Institute)
Considerati/ECP's Artificial Intelligence Impact Assessment (AIIA)

While the difficulties of developing an AI risk assessment framework have been well documented,[IX] a thorough and consensus-supported AI risk framework remains elusive.

Automated Decision Impact Assessment (ADIA): A Potential Path Forward

In Facebook's response to the European Commission White Paper on AI, Facebook stressed the need to align with GDPR around self-assessment of AI risk, advocating that any new AI regulation should build upon the requirements that already exist in GDPR in order to provide greater legal clarity, avoid duplicative regulation, and ensure a proportionate approach to these novel issues. Through GDPR, the Commission established the duty to implement accountable data protection programmes that include Data Protection Impact Assessments (DPIAs): ex ante self-assessments for data processing likely to be high-risk. Based on that model, Facebook suggested the possibility of a similar approach to AI, developing the concept of Automated Decision-making Impact Assessments or ADIAs,[X] akin to DPIAs, as a way to assess, determine, and document the level of risk posed by AI applications, and to mitigate those risks accordingly.[XI]

In its response, Facebook described how the basic elements of an ADIA process could be codified in regulation, complemented by evolving soft law instruments that would provide more detailed guidance.[XII] This guidance could include a detailed taxonomy of the kinds of risks and harms to be considered, indicative examples of AI uses that are presumed to be high-risk (a presumption that could be rebutted with appropriate mitigations documented in an ADIA), and a step-by-step methodology that developers could follow when seeking to identify and quantify harms. Such guidance would help ensure adequate consideration of the specific context of the automated decision-making at issue, and a greater focus on concrete and measurable harms, while also ensuring consideration of the application's benefits as well.

In terms of embedding such an ADIA framework into a system of regulatory enforcement, Facebook suggested that enforcement actions could be triggered when companies fail to properly conduct a risk assessment or reasonably mitigate the risks they identify. Consistent with GDPR's approach, a prior consultation with the relevant regulator would be required only when the ADIA process has resulted in the identification of residual high risks for which appropriate mitigations are not reasonably available or have not been identified. This would encourage organisations to proactively consider and adopt mitigations that reduce the initial high risk to an acceptably low level.

Facebook offered this potential ADIA framework as a more flexible alternative to the EC White Paper's proposed system of enforcement, which would require prior conformity assessments of AI systems by regulators or third-party auditors before those systems are deployed in the EU. As articulated in Facebook's response, the enforcement system proposed by the Commission could risk unnecessarily overburdening AI developers and significantly impairing innovation and economic growth that would benefit European citizens.

Prototyping an ADIA Framework

Following up on Facebook's description of a potential ADIA framework, the company partnered with a group of ten other European AI companies under the Open Loop initiative to co-create a draft ADIA framework (policy prototype), and to test it in practice by running that AI risk assessment process on a selected set of real world AI applications. The consulting firm Considerati contributed to the methodology, content and analyses of this policy prototyping program.

Based on Facebook's initial conceptualization of an ADIA framework, composed of a legally codified ADIA requirement complemented with more detailed guidance via soft law instruments, Open Loop members – including Facebook and Considerati – co-created a prototype ADIA framework and structured it into two parts: an ADIA prototype law, which was drafted as a legislative document with articles and recitals; and an ADIA playbook, which provided comprehensive guidance aimed at helping companies interpret and comply with the legal text and conduct their ADIAs. The playbook included a step-by-step risk assessment methodology; a list of values relevant to AI/ML and ADM; a taxonomy of harms; and examples of mitigating measures.

Through a policy prototyping methodology (as explained below), we then tested the ADIA framework as normative guidance provided directly to AI developers. The focus of the testing was twofold:

- evaluate the ADIA framework and understand its applicability, feasibility, merits and limitations within the reality of corporate practices, and across a broad and diversified range of AI applications;
- recommend specific improvements to the draft ADIA framework, and derive evidence-based policy recommendations to inform ongoing AI regulatory discussions.

It is important to note that the ADIA framework that we prototyped is not itself a legislative proposal, but an instrument aimed at exploring and testing alternative policy frameworks and regulatory pathways. The ADIA framework was developed solely for the purpose of being tested and experimented on as a possible option for how AI-related risks could be identified, assessed, and mitigated. It is a departure point, a platform for experimentation, and not a final conclusion.

Beyond testing and evaluating the merits and limits of this governance framework, this project also gave us an opportunity to test the process of policy prototyping itself as a sound methodology to inform rule-making activities.

The project consisted of:

1. **Co-developing an ADIA framework**
2. **Testing and evaluating it with a group of selected AI companies**
3. **Revising the draft framework**

and

4. **Delivering policy recommendations based on that empirical testing and evaluation, in order to inform the evolving AI governance debate.**

Policy prototyping

02

Policy prototyping

What is policy prototyping?

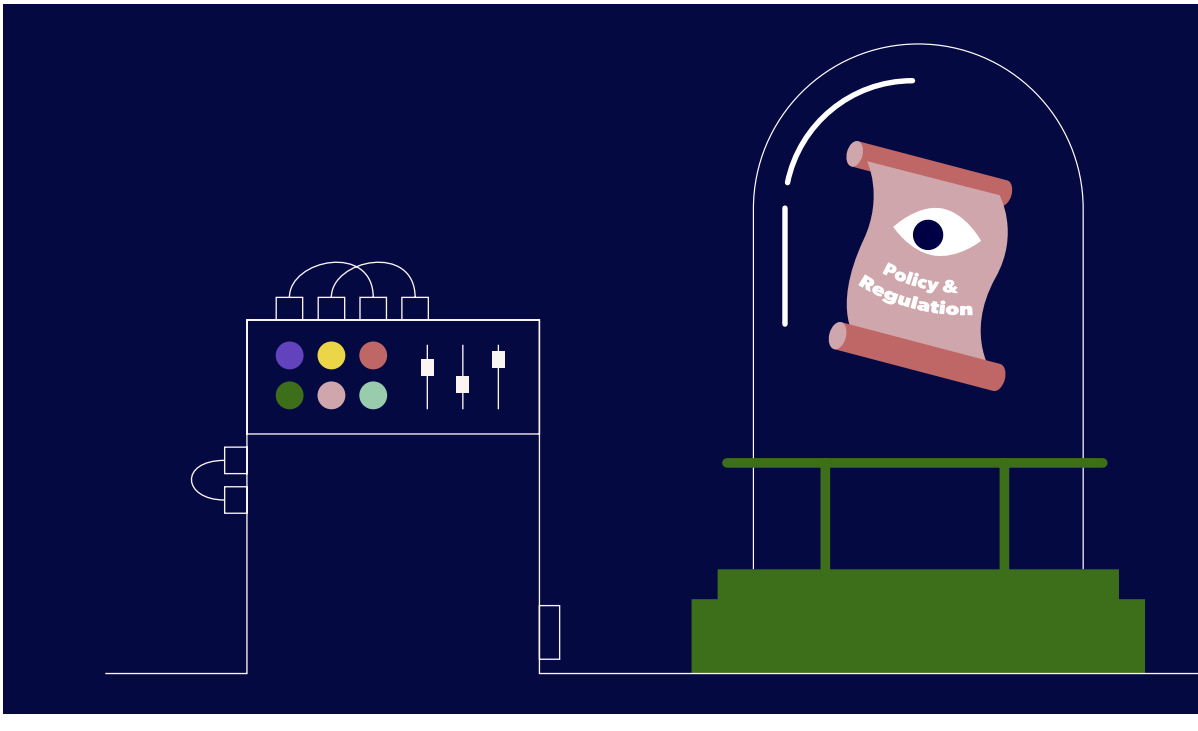
Policy prototyping is a methodology to test the efficacy of a policy by first implementing it in a controlled environment. Policy prototyping applies a user-centered design and user research approach, which is commonplace in product and service design, to the development of law and policy.[8]

Legal philosopher Lon Fuller has defined law as the enterprise of subjecting human conduct to the governance of rules.[9] Rules (policies) are made to influence the behaviour of individuals, groups or organisations (the norm addressees) with the goal of bringing about certain mutual behaviour, action or abstention from action.[10]

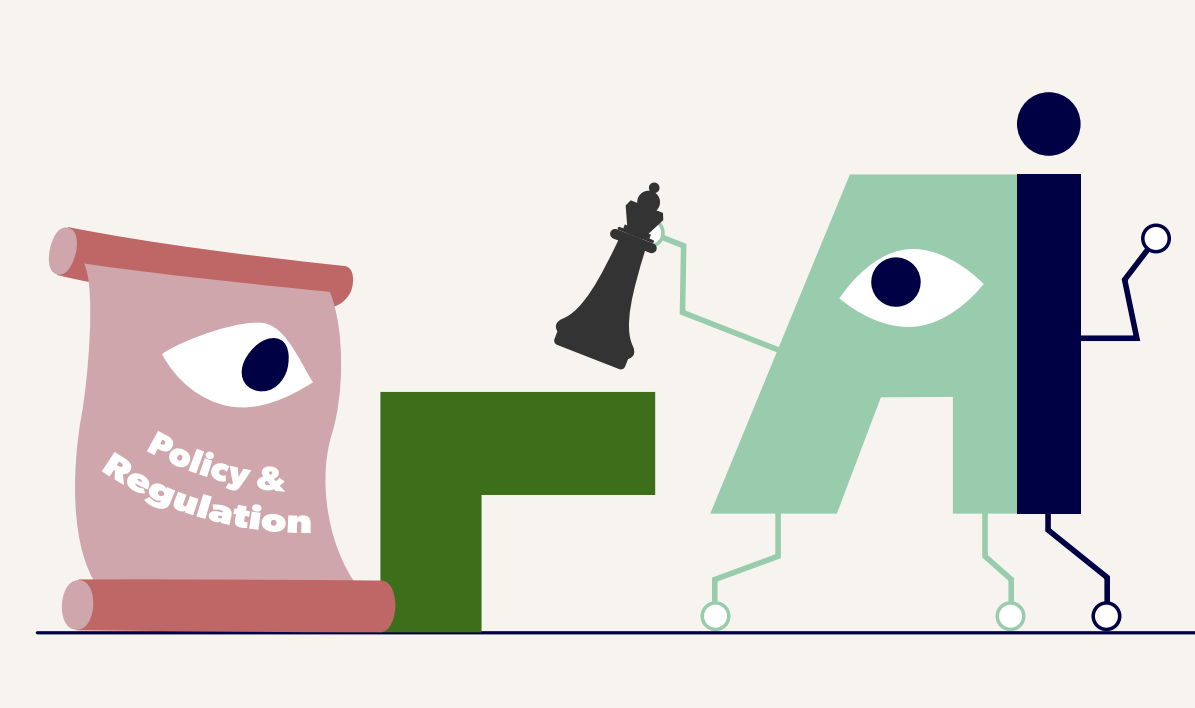
In other words, a law or a policy – as instruments directed at producing certain effects – is a means to achieve a particular policy goal.

Nonetheless, it is difficult to know (and anticipate) the effects produced by laws before they are enacted and put into force. This is particularly true with laws governing new and emerging technologies. And although proposed laws and policies are often discussed and debated extensively, they are seldom tested in practice.[11] As such, laws are typically enacted without it being clear whether they actually will be effective and 'fit for purpose'.

In this particular policy prototyping program, we wanted to test whether an ADIA approach to AI policy would be effective in achieving its intended AI governance goals. We did this by first creating a prototype law: a normative framework built for the sole purposes of being tested by a limited group of norm addressees, and aimed at producing actionable feedback and concrete policy recommendations to inform rule and law-making processes.



8. See Brown and Katz 2011; Villa Alvarez, Auricchio, and Mortati 2020; Kontschieder 2018. See also Brown 2008.
9. Fuller 1964.
10. Kelsen 1941.
11. See, e.g. Bason 2016.

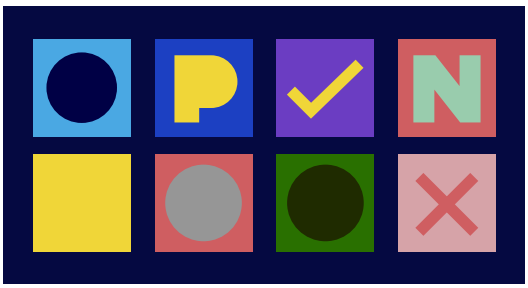


Why policy prototyping?

The idea is that policy prototyping will lead to more effective and evidence-based policy-making and avoid the societal costs of 'bad policy'. These costs can be of an economic nature (e.g. high compliance costs, high enforcement costs, or loss of opportunity), infringements of rights and freedoms, or unintended consequences and collateral effects.

Policy prototyping may be especially useful in areas where the pace of technological development and innovation is high and where formal legislation tends to struggle to keep up. Prior to rolling out a new governance framework (proposed law, codes of conduct, standards, guidelines, etc.), policy prototyping can be a swift and agile way to understand that framework's effects, strengths and limitations. In design thinking, a prototype is "the visible, tangible or functional manifestation of an idea, which you test with others and learn from at an early stage of the development process." [12]

A prototype can thus be seen as a low-resource, more quickly deployed version of an idea, used to run experiments in order to test that idea and learn whether to pursue and invest in it more fully. This is similar to beta-testing cycles in technology, a 'trial and learn' process that informs the final tool or application.



Policy prototypes can help the makers and users of policy better understand the extent to which that policy is clear, relevant and effective before turning it into a more robust, fully fleshed out version that is ready to be released and applied more broadly.

12. Leurs and Duggan 2018.

The EU ADIA policy prototyping program

03

The EU ADIA policy prototyping program

AI risk assessment

We chose the topic of AI risk assessment as the focus for our policy prototyping program. As described in the introduction, AI risk assessment features prominently in the debate on AI governance and is mentioned in different policy responses to the potential risks of AI.

For instance, the proposal for an Algorithmic Accountability Act in the United States is specifically aimed at introducing an "automated decision-making system impact assessment."[13]

In the EU, lawmakers are following a risk-based approach and contemplate making distinctions between low and high-risk applications.[14] Finally, the Council of Europe recommends the introduction of human rights impact assessments for AI as a precautionary measure.[15]

The topic of risk assessment for AI/ADM lends itself particularly well to a policy prototyping exercise as there are no individuals involved that may suffer harm as a result of the prototype. Our prototype law only sets requirements that affect the participants in our exercise (i.e. the developers and users of AI), not those who are affected by the applications (e.g. patients, citizens, consumers).

Project overview

The EU ADIA policy prototyping program was designed to contribute practical insights to the current policy debate on AI impact assessments. To this end, we designed a prototyping method suitable for a four week program, and drafted a prototype law and supporting documentation.

To test the prototype in a real-life setting, we selected 10 European AI startups* in various sectors willing to join the program to implement the prototype and share their experiences. To enable those organisations to fully participate in the risk assessment, we ensured that no disclosure of proprietary or sensitive information was needed and that the program would not result in a value judgement about their products or organisations. The program aimed to evaluate policy, not products or services.

These organisations operate in a wide range of sectors from healthcare to financial services and consumer applications. Some organisations provide platforms or develop solutions for implementation by others, others have a B2C business model. The participating companies were asked to select an AI application that would produce effects or have an impact on people, and simulate the application of the ADIA process on that particular application.[16]

* Based in or holding key operations in Europe/across EU

13. H.R.2231 – Algorithmic Accountability Act of 2019.
14. See EC 2020a.
15. Council of Europe 2020.
16. This was done at the onboarding phase of the program via an initial sign-up form, in which they were asked: 'You will test the ADIA based on an existing AI application in your company that you can self-select (i.e. products or services powered by AI/ML). Please tell us about this AI application you would like to run the impact assessment on.'

ADIA Policy Prototyping Program participants

Allegro.ai (IL) provides an AI/ML software infrastructure (open-source / enterprise) helping companies develop machine- and deep-learning products. Allegro.ai ran the impact assessment on this AI/ML software infrastructure (open-source / enterprise) as it helps companies build fully automated learning pipelines with automatic feedback loops, allowing for a fully autonomous decision-making system.

RiAtlas (IT) is a digital healthcare startup developing solutions such as remote monitoring and a smart patient health classification. This company applied the ADIA framework to their core AI application which, supported by machine learning and predictive models and structured on "validated" clinical datasets, classifies patient health status from clinical and personal data collected (patient-reported outcomes from a mobile app and vital signs from a smartwatch). This tool supports clinical decision-making tasks such as patient's health status classification, smart data visualization and early detection of clinical risks.

NAIX Technology (DE) ran the risk assessment on their core application and service. Based on AI and natural language processing (NLP), NAIX Technology developed a software to automatically anonymise or pseudonymise personal identifiable information (PII) in large sets of documents, helping companies meet the requirements of GDPR.

Evo Pricing (UK) develops an Autonomous Supply Chain solution of price management, promotion, forecasting and supply decisions. The AI application they selected uses big data to increase the efficiency of supply chain decisions, which in turn helps reduce waste, increase market efficiency, enhance product availability and service levels.

Keepler Data Tech (ES) is a software company specializing in the design, construction and operation of data products based on public cloud platforms. The AI application that Keepler ran the ADIA on automates the task of assigning service responsibility and extracting important content from insurance claims. Within the context of an email customer service management project, the application reduces the manual workload while enhancing focus time on content as opposed to processing documents.

Unbabel and Reface's participation was limited to the first week of the program.

Unbabel (PT) provides an AI and human driven translation-as-a-service platform for enterprise clients. Their product enables enterprise customers to provide multilingual customer support to their users by removing language barriers. The AI enabling technology is Machine Translation, augmented with human editing in some use-cases.

Feedzai (US) provides a risk management platform to prevent financial crime. They performed the AI risk assessment process on an application that automatically determines the fraud risk of new bank account opening applications. The system has access to demographic data (filled in by applicants) and, based on the predictive risk of fraud, supports decisions around providing or denying people access to banking services.

Rogervoice (FR) conducted the AI risk assessment on an application that allows deaf people to make calls by using Speech to text – Text to speech technology. This AI application makes communication via telephone accessible to a sector of society that previously could not make use of this service, and allows it to cross the boundaries of communication between people.

Irida Labs (GR) is an edge AI and computer vision software company with a mission to bring vision intelligence to any device. They ran the ADIA framework on their end-to-end AI software, which integrates ML detection models for people, vehicles and other objects with vision system design and data management processes. Their tools empower the development of vision-based solutions for Smart Cities, Smart Retail, Industry 4.0, Surveillance and Logistics. Examples of applications powered by Irida labs technology include smart retail analytics (customer count, customer engagement analytics, waiting times and queue flows analytics), free flow vehicle monitoring, parking space management (occupancy monitoring, zone management), and process automation in warehouses and construction sites.

Reface (UA) develops an AI-driven face-swapping application that enables to transpose faces in photos and videos. Their tool for hyper realistic face swapping was the one used for testing the ADIA prototype.

The participant organisations have different levels of familiarity with translating legal requirements to their own products and processes. More than half of participants have experience with performing risk assessments, mainly data protection impact assessments (DPIAs, a GDPR requirement).

Methodology

In this program we set out to test the effectiveness of our policy prototype in achieving a predetermined policy goal, which is described in detail in chapter 4.

The prototype, which was drafted in the format of a law, was tested against the following three criteria:

1

Policy understanding

A critical necessity for any law or policy is that the norm addressee – those who are subject to the policy requirements – actually understands what is required of them. As such, the clarity of a policy or law is of vital importance, not only from the perspective of compliance, but also from the perspective of legal certainty.[17]

2

Policy effectiveness

A law or policy should obviously contribute to reaching the overall policy goal. If the policy does not substantially contribute to the achievement of the policy goal, it should not be implemented. The policy goal for the Automated Decision Impact Assessment (ADIA) prototyping program is described in chapter 4.

3

Policy costs

Compliance with a law or policy may entail certain costs. These can be simply the cost of compliance or oversight (i.e. what resources did it cost to comply or ensure compliance), but may also include the costs of unintended side effects of the policy (e.g. negative impact on innovation or infringement of human rights). A policy can only be considered successful when the importance of reaching the policy goal outweighs the costs associated with reaching that goal through the implementation of the policy. Note that costs may be distributed unevenly over stakeholders. Those who actually bear the costs associated with the implementation of a policy, and whether the costs to them are fair, should also be taken into account when assessing the success of a policy.

17. In law, this principle is referred to as lex certa.

ADIA Policy Prototyping program timeline

Kick off co-creation workshop

Week 1
Introduce and understand AIDA

Inquire exposure to risk/ impact assessment and get to know prototype

Week 2
Implement ADIA

Simulate ADIA, i.e. implement and quantify risks on self-chosen application

Week 3
Reflect ADIA implementation

Review week 2 activity on implementing and quantifying risks

Week 4
Implement with playbook and reflect

Probe guiding material to ADIA provisions and analyse before-after

Closing co-creation workshop

Research approach

In week 1, participants were introduced to the prototype law and were asked about their understanding of key concepts and terms of the prototype law, along with their prior experience with impact assessments.

In week 2, we asked the participants to simulate the implementation of the ADIA and document the outcomes.

Participants were asked to reflect on their experiences implementing the prototype law in week 3. We asked them to comment on the clarity of the requirements, whether they could apply them to their own context, and if they believed the requirements were useful.

In week 4 the users were presented with a 'playbook'. This playbook provided participants with additional guidance on procedural and substantive aspects of performing the ADIA through:

- A step-by-step risk assessment methodology;
- An overview of values often associated with AI applications;
- A taxonomy of harms;
- Examples of mitigating measures.

Elements from the playbook were sourced from publications in the AI/digital ethics domain, authored by:

- Governmental and political actors at state, national, and supra-national level;[18]
- Competent supervisory authorities (data protection authorities, consumer protection authorities);[19]
- International organisations (e.g. UN, OECD, Council of Europe);[20]
- Industry groups and professional associations;[21]
- Consumer advocacy and civil rights groups;[22] and
- NGOs and think tanks.

The playbook simulated the common occurrence of dissemination of additional guidance by supervisory authorities after a norm-based policy comes into force.

A closing co-creation workshop marked the end of the four core testing weeks with the ten participating companies. The purpose of the workshop was to present the preliminary findings from the program, to follow up on specific themes that emerged from the feedback collected,[23] and to share and further discuss these findings and themes with relevant audiences, notably with the participating companies, EU Institutions and EU Member State policy representatives, and academics and industry peers with experience and/or interest in policy experimentation.

18. See, e.g. EC 2019a and EC 2019b.
19. See, e.g. Information Commissioner's Office (ICO) 2017. See also EC 2017.
20. See, e.g. OECD 2019, Council of Europe 2019.
21. See, e.g. IEEE 2017.
22. See, e.g. UNI Global Union 2018.
23. Reisman et al. 2018.

Data collection

Data was collected via regular surveys sent to the participants each week. We used a mobile ethnography approach to collect data from participants.[24]

This was an innovative methodology: while gathering feedback from users in this way is common in the world of product and service design, it has to our knowledge never been done in the field of law and policy.

We used a smartphone application to gather feedback from participants.

Feedback consisted of answers to multiple choice/Likert scale questions, as well as free format responses, for instance in the form of text, audio, video, mind maps and flowcharts.

We collected further direct feedback on the prototype law from both participants in the exercise and other stakeholders by keeping the prototype law open for comments and edits throughout the program.

Limitations of the exercise

While this prototyping exercise gives valuable insights into both the effectiveness of a policy prototyping exercise in general, and specific insights into a policy for mandating automated decision-making impact assessments, it has some limitations.

First of all, there were a limited number of participants. As such, gathering representative quantitative results was not possible. For the purpose of this prototyping exercise, however, this was not a significant problem as we mainly wanted to collect qualitative feedback on our prototype. The qualitative approach we took, and in particular the mobile ethnography element of our approach, allowed us to observe the companies' process of performing the ADIA in the context of their business models, everyday operations and real world settings, and applied to their own AI applications.[XIII] Our focus on qualitative measures was consistent with calls by some stakeholders for a greater focus on qualitative dimensions in impact assessment procedures for the identification of complex and uncertain risks, like those potentially posed by AI systems.[XIV]

The second limitation was the available time for the prototyping exercise. Depending on the complexity of the risk assessment framework and the requirements captured therein, a comprehensive risk assessment exercise may take anywhere from three months to over a year. As this project was programmed to last 6 weeks in total, the participants had limited time to do a full ADIA-based on the prototype law requirements. The time limitation was embedded by design into the program, and inspired by the use of "sprints"[25] and other "agile" approaches to policy making, namely when it comes to new and emerging technologies.[26] The participants were thus asked to simulate the application of the ADIA process, which meant doing a more 'high-level' assessment without the requirement of documenting their results in full detail as is normally required in impact assessments. As such, answers to some of the questions (such as the time necessary for doing a full scale ADIA) are estimates rather than concrete numbers.

A third limitation had to do with the enforcement element of the program. Given the nature of this exercise, the requirements that we asked participants to follow and implement were not rigorously enforced. This policy prototyping program was not focused on enforcing compliance with the requirements laid out, but about documenting

24. www.dscout.com
25. Kimbell 2015.
26. World Economic Forum (WEF) 2018.

compliance with those requirements and assessing its understanding, effectiveness and costs. In future iterations of this kind of prototyping program, it would be most useful to have regulators join these efforts to help test the effective enforcement of this type of prototypical framework.

A final limitation is the diversity of the participants. While the participants come from different countries, have diverse cultural backgrounds and, most importantly, develop different applications and operate in different sectors according to distinct business models, they were all in the startup or scale up phase of their organisation. As such, the results of this prototyping exercise are not necessarily representative for medium-sized enterprises or multinational organisations. However, our priority was to ensure the gathering of actionable feedback from companies that, generally speaking, would not have the resources that other larger organisations would. In other words, we wanted to ensure that the requirements of ADIA frameworks could be complied with by companies of smaller size.

The Automated Decision Impact Assessment (ADIA) policy prototype

04

The Automated Decision Impact Assessment (ADIA) policy prototype

Policy goal

In this chapter, we describe the Automated Decision Impact Assessment (ADIA) policy prototype (referred to as "the prototype law").

In its Ethical Guidelines for Trustworthy AI, the European Union sets out the requirement that AI must be trustworthy:

"Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems. Without AI systems – and the human beings behind them – being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered, preventing the realisation of the potentially vast social and economic benefits." [27]

In order for AI or automated decision-making to be trustworthy[32], its application should be 1) legitimate, 2) ethical, and 3) robust.

To determine whether automated decision-making systems are indeed legitimate, ethical and robust, we must first establish what the potential unwanted consequences of AI/ADM could be and how they might affect the rights and freedoms of persons or groups. In particular, this task falls to the developers and users of ADM systems.

27. EC 2019a.

As such, our overall policy goal is to:

Make developers and users of ADM systems aware of the risks their applications may pose, and make them find ways of mitigating these potential risks.

Developers and users conduct effective risks assessments for their ADM application.

1. **Users are able to identify what risks their applications may entail for the rights and freedoms of subjects**
2. **Users are able to determine how significant these risks are (e.g. high or low)**
3. **Users are able to formulate mitigating measures to these risks**
4. **Users are able to adequately assess whether these measures remove the risks or reduce them to an acceptable level (residual risk)**

For this policy, our 'Theory of Change' can be shown as follows:[28]

Policy evaluation

Policy content
Is the policy sufficiently clear (i.e. can the norm-addressees understand the content)?

Policy implementation
Were the participants able to implement the policy? What was the effect on them (e.g. in terms of resource requirement)?

Policy impact
Didi the policy achieve the desired effects? Were there any unintended consequences (both positive and negative)?

Policy goal:
Make developers and users of automated decision-making systems aware of the risks their applications may pose, and make them find ways of mitigating these potential risks.

Outcome:
Developers and users conduct effective risk assessments for their ADM applications.

Requirement 1:
Users are able to identify what risks their applications may entail for the rights and freedoms of subjects. (Art. 4.1)

Requirement 2:
Users are able determine how significant these risks are (e.g. high or low). (Art. 4.2-3)

Requirement 3:
Users are able to formulate mitigating measures to these risks. (Art. 4.4)

Requirement 4:
Users are able to adequately assess whether these measures remove the risks or reduce them to an acceptable level (residual risk). (Art. 4.5)

28. A 'theory of change' defines long-term goals and then maps backward from those goals in order to identify necessary preconditions. See Brest 2010. See also www.theoryofchange.org

The prototype law and its requirements

Based on the policy goal and our associated requirements, we drafted a prototype law. The prototype law was formulated and structured in the same style as an actual law.[29]

The point of departure for the prototype law was that it be technology neutral (hence the use of automated decision-making rather than AI/ML in the text) and principle-based. In this way, the prototype law could be applied to different technologies, sectors and contexts.

Recitals	Content
Art. 1 subject matter and objectives	Defines the objective of the prototype law: protection of fundamental rights and freedoms, ensuring a trustworthy application of ADM, stimulate development and use of ADM for well-being of society.
Art. 2 material scope	Sets the scope for the prototype law to development, production, distribution and use of automated decision-making systems whose use may have a significant effect on natural and legal persons
Art. 3 definitions	Defines the actors relevant to the prototype law and concepts related to automated decision-making system.
Art. 4 risk assessment	Sets requirements for the performance, timing, and contents of an ADIA. Defines when ADIA outcomes warrant prior consultation with the supervisory authority.

Regarding Art. 4 in particular, it is important to note that the prototype law mandates that the risk assessment process:

- should be conducted prior to the deployment of the automated decision-making system, and in cases where the application of the ADM system is likely to result in a high risk to rights and freedoms of natural and legal persons, namely in cases of:

- potential unfair bias or discrimination;
- potential loss of control or agency for the subject, including economic or psychological manipulation;
- large scale application of automated decision-making.

– should contain at least:

- a detailed description of the automated decision-making system, its design, its training, its data, and its purpose;
- an assessment of the quality, integrity and representativeness of the data used to train the underlying model;
- an assessment of the risks involved for natural and legal persons, with a specific focus on subjects and for end-users; and,
- the measures envisaged to address the risks.

29. The full prototype law can be found in annex 1.

ADIA prototype policy evaluation

05

ADIA prototype policy evaluation

As discussed in chapter 3, we tested the prototype law against the following criteria:

- Policy understanding
- Policy effectiveness
- Policy costs

Assessment of policy understanding

Policy understanding is essential for an adequate implementation of the prototype law and thus for achieving the policy goals. As the prototype law is designed in a norm-based and technologically neutral manner, norm addresses need to be able to understand the concepts, norms, and requirements and be able to apply these to their own context and situation. This enables us to assess what is needed to increase policy effectiveness and to demonstrate the difference between the clarity of the policy in theory and in practice.

We tested the understanding of the prototype law contents by focusing on a set of key definitions, namely the definition of: 1) an automated decision-making system, 2) high risk, and 3) the actors involved. We also tested participants' understanding of the requirements for doing an ADIA specified in article 4.4 of the prototype law.

Definitions (Arts. 2-3)

'Automated decision-making system' means a computational process derived from machine learning, statistics, artificial intelligence or other data processing technique, that makes a decision or facilitates human decision making. (Art.3e)

Automated decision-making system

The concept of an 'automated decision-making system' is not unique to the ADIA prototype. 'Automated decision-making' is used in existing legislation such as the GDPR (Art. 15 and 22) and corresponding guidance by supervisory authorities, as well as in the proposed Algorithmic Accountability Act. The concept of an 'automated decision-making system' is central to the prototype as it is the object of the risk assessment, so it's critical that it be understood.

Participants were divided on the clarity of this concept. They appreciate how this definition applies to both autonomous and non-autonomous systems and is not scoped around a particular technique (from rule-based systems to deep learning). For some, however, it was unclear how much of the decision-making process was in scope (i.e. is a decision by a human supported by the system part of the 'system'?). Others wondered what the limits of 'facilitates' are, (i.e. should data extraction

and visualization be considered an 'automated decision-making system'?). As noted by one of the participants, *"a software that visualizes data can for instance fall under the description – a computational process that uses a data processing technique (data visualization) that facilitates human decision"*. In this manner, the definition may be read much more broadly than policymakers might intend.

On the other hand, during the co-creation session, the point was raised that 'decision-making system' may in some ways be too *narrow* a definition.

Many AI applications support humans with operations that do not necessarily equate to making decisions (e.g. text mining, speech-to-text applications, or eDiscovery). These applications may not be considered decision-making systems in the narrow sense of the word, but nevertheless can have an impact on the rights and freedoms of data subjects. Although the definition in the prototype law itself is likely broad enough to capture those AI/ML applications of concern to policymakers, the specific term 'automated decision-making system' may still confuse norm addressees.

High-risk

Whether there is a high risk to the rights and freedoms of natural and legal persons must be judged on the context, nature, purpose and scope of the application. There is a high risk when there is a significant chance that the automated decisions made by the automated decision-making system, or the subsequent actions taken by users, end-users or subjects on the basis of that automated decision, result in negative effects with a significant adverse impact on the rights and freedoms of natural and legal persons. recital 11

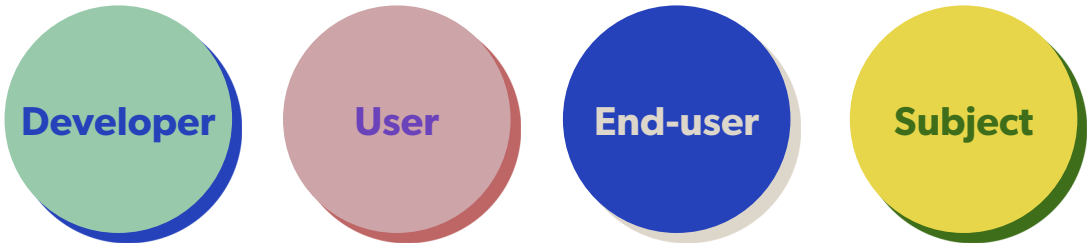
All participants felt they had a basic understanding of the definition of 'high-risk'. When actually assessing risk, however, most users seemed to focus mainly on the functional risks of their application – that is, risks related to how their AI systems are built and how they operate, such as potential bias in data sets, concept drift or model performance. Risks related to broader structural aspects – such as concerns related to the ethical application of automated decision-making systems and the consequences of these decisions (such as impact in terms of fairness, proportionality, societal effects) – were given less attention. [30] In effect, participants understood and applied a much narrower definition of "high risk" than what was actually articulated in the prototype law. For more information, see the discussion on the ADIA requirements.

30. The categorization of functional and structural risks is based on the distinction between of epistemic and normative concerns elaborated by Mittelstadt, B. D., Allo, P. Taddeo M. R., Wachter, S., Floridi, L. (2016), The ethics of algorithms: Mapping the debate, in: Big Data & Society July-December 2016.

Actors

The prototype law defines four actors relevant to ADM systems:

- **'Developer' means the natural or legal person responsible for the technical development of the automated decision-making system.**
- **'User' means the natural or legal person deploying an automated decision-making system to achieve a particular goal.**
- **'End-user' means the natural or legal person using the automated decision-making system for the purposes intended by the user.**
- **'Subject' means the natural or legal person subjected directly or indirectly to a decision of an automated decision-making system. (Art.3f-i)**



The prototype law defined 4 different types of actors and applied these concepts from an organisational (and independently operating individuals) perspective, and based on their appropriateness and capability to undertake risk assessment and mitigation tasks. In other words, this taxonomy designates these roles by taking into account which actors would be best positioned to identify, assess and mitigate risks within the AI product/service lifecycle. It is important to note that these actors are non-exclusive, and that organisations developing and deploying their own technology (most participants) are both 'Developer' and 'User'.

For instance, Reface develops and deploys their own system for face-swapping in photos making them both 'developer' and 'user'.

Although the definitions of actors were considered mostly clear and useful by almost all participants, we received important feedback about a mismatch between how these terms are used in the ADIA framework and how they are used in practice, which may trigger confusion about how some roles were not covered by the proposed taxonomy; and about the need to take into account the specific characteristics of some AI products and services when deciding who should conduct the risk assessment.

Firstly, the meaning of the terms to designate the actors in the prototype law is very different from the day-to-day use of these terms. This is particularly true with GDPR, where the parties corresponding to ADIA "users" would be called 'subjects'. The term 'developers' in the ADIA framework may also consist of internal teams within the same company, and this will often overlap with the separate definition of the term 'user'. The different uses of these terms was deemed to be potentially confusing.

Secondly, and although multiple 'actors' can refer to the same organisation, this does not cover all possible roles. Allegro.ai, for instance, provides a platform for managing machine learning/deep learning project lifecycles.

31. For the purpose of the policy prototyping exercise, Allegro assessed the risk of a hypothetical AI vision application that a customer would run on their platform.

While this seems to fall more in the category of "developer", it is unclear where this category fits in practice. As Allegro.ai stated:

"Since we are an AI infrastructure company, these sections do not relate to our product directly. That said, we will assume we provide a service on top of our platform, that does continuous training for a specific computer-vision task, and try to answer the questions in this context." [31]

Allegro.ai consider themselves "provid[ing] infrastructure for developing, deploying and monitoring such [automated decision-making] systems."

"[They] do not directly build them,", they stated. *"It does not apply to our application, we are building the infrastructure for such ML DL applications,"* they clarified on several occasions. This implies that a platform provider enabling AI or ML applications might not understand themselves as a 'developer' or 'user' and hence the prototype law may require additional actor categories that either take into consideration a more ample value chain, or that assign actor categories for specific automated decision-making use cases/business models from the outset.

Thirdly, it is important to understand how the AI application being built will actually contribute to powering automated decisions. Here, the difference between providing a generic learning algorithm and a pre-trained off the shelf model is relevant. While the providers of learning algorithms have no control over the training data and the eventual application of the model, the actions of a provider of pre-trained models will have a more direct relationship to specific downstream decision tasks. A better understanding of the type of AI application being built should inform the development of future taxonomies of AI actor roles, along with the proper assignment of responsibility for conducting risk assessments.

In sum, the taxonomy of actors and roles in the prototype law is a good starting point and its granularity is aligned with the multiplicity of actors involved in the AI ecosystem and the complexity of their interactions. But there is work to be done in refining this taxonomy and adapting it to how these terms are used on a daily basis, and expanding it to cover additional types of AI actors and AI applications. In practice, and based on the feedback of participants, the 'AI value chain' and the interaction between 'developers' and 'users' is quite complex, which makes it harder to establish who is responsible for assessing the risk, who is best equipped to assess the risk, and who should mitigate the risk.

To increase the clarity and applicability of this section of the prototype, the description of actors in the prototype should likely be revised.

Risk assessment (Art. 4)

Risks which in any case require an ADIA (Art 4.3)

An automated decision impact assessment referred to in paragraph 2 shall in any case be required in case of:

- **potential unfair bias or discrimination towards subjects, including price discrimination, employment discrimination or unfair differential access to services;**
- **potential loss of control or agency for the subject, including economic or psychological manipulation;**
- **large scale application of automated decision-making, including profiling and systematic monitoring, that may affect communities or society as a whole.**

When asked to imagine how these risks apply to their applications, participants demonstrated that they were able to identify ways that these risks could manifest themselves through their applications. For example:

- Feedzai, which assessed its application for detecting fraudulent bank account openings, identified a risk for unfair differential access to financial services through disparities in false positive rates in their application based on specific sensitive attributes (such as age, gender, employment status, zip code, or income).
- Reface identified a risk for psychological manipulation if their 'face swap' application is used for the generation of misleading content.
- Unbabel identified a risk for potential loss of control if their translation system makes a 'polarity error', for instance erroneously translating "do" instead of "do not."

However, there was some variation in how concepts such as "unfair bias" and "loss of control or agency" were perceived. For instance, bias was mostly understood as bias in the data. However, Feedzai identified a potential bias that went beyond its data, noting that unbanked customers are unable to access credit and are also not able to demonstrate creditworthiness. This example demonstrates how some participants were able to apply these concepts beyond technical and functional considerations of the ML models behind their AI applications, although most others focused their assessments within a model-centric understanding of their applications. This means that risks related to the overall socio-technical system in which an AI application lives and behaves (which may include other software components, data sources, and interfaces) are much less likely to be recognized.

The concept of "large scale application" in the definition was also unclear to the participants. There were different perspectives on what should be considered and when that constitutes a large scale. Some participants understood this concept in terms of:

- model characteristics (volume or number of elements in a model) and volume of data;

"Large scale is different in every vertical. It does not apply to our solution, but we define 'large scale' for customers having more than one million annotated entries." (Allegro.AI)

"Our data processing is definitely 'large scale'. Translation models are trained on millions and billions of words and sentences… and consume significant resources. They are then applied to translation of billions of words for our customers. Scale here has to do with the number of "atomic" units (words and sentences) needed to train the models" (Unbabel)

- the sector where the application is applied (a niche market is considered small scale);
- the different attributes and characteristics of end-users (languages, age groups);

"Being a technology that is used in general and not in a specific environment, the challenge is to be able to cover all the necessary aspects to ensure a good translation / interpretation service. In other words, consider gender, languages, accents, ages, etc." (Rogervoice)

- the number of subjects and duration of processing.[xv]

"Irida Labs' AI processing could be considered as large scale data processing, since there might be applications involving

large numbers of subjects and prolonged duration of processing (i.e. surveillance camera on a shopping mall's entrance recording 24/7)."
(Irida Labs)

Additionally, some participants thought that more categories of cases requiring an ADIA should be added (for instance, cases where there is potential for misuse, privacy violations, or feedback loops that reinforce structural biases).[XVI]

Some participants also misunderstood the meaning of this article, interpreting the specific examples in Art. 4.3 to be the *only* risks for which an ADIA had to be performed, highlighting the need to clarify in the text that the list of included examples is not exhaustive.

Minimal requirements of an ADIA (Art.4.4)

4.1 **Prior to the deployment of an automated decision-making system, the user shall assess the risks of the envisaged automated decision-making system and its application on the rights and freedoms of natural and legal persons.**

4.2 **In those cases where the application of an automated decision-making system is likely to result in a high risk to rights and freedoms of natural or legal persons, the user shall carry out an automated decision impact assessment prior to the deployment.**

4.3 **An automated decision impact assessment referred to in paragraph 2 shall in any case be required in case of:**

- **potential unfair bias or discrimination towards subjects, including price discrimination, employment discrimination or unfair differential access to services;**
- **potential loss of control or agency for the subject, including economic or psychological manipulation;**
- **large scale application of automated decision-making, including profiling and systematic monitoring, that may affect communities or society as a whole.**

4.4 **An automated decision-making system impact assessment shall contain at least:**

- **a detailed desciption of the automated decision-making system, its design, its training, its data, and its purpose;**
- **an assessment of the quality, integrity and representativeness of the data used to train the underlying model;**
- **an assessment of the risks involved for natural and legal persons, with a specific focus on subjects and for end-users; and**
- **the measures envisaged to address the risks including safeguards, security measures and mechanisms protecting the rights and freedoms of end-users and subjects, and to demonstrate compliance with this Policy Prototype, taking into account the rights and legitimate interests of those concerned.**

4.5 **In those cases where the automated decision impact assessment indicates that the application may result in a high risk to the natural rights and freedoms of natural and legal persons and these risks can or will not be mitigated, the user shall prior to the deployment consult with the supervisory authority.**

Article 4.4 sets the minimal requirements for an ADIA. Most of the participants considered these elements to be useful in general. Some argued for additional requirements. One idea was for an assessment of end-user understanding and capabilities, to ensure that automated decisions are not misunderstood or mis-implemented. Another suggestion was an assessment of whether the application is necessary and proportional, and if there are other less risky solutions possible.

a detailed description of the automated decision-making system, its design, its training, its data, and its purpose;

This requirement was clear enough for almost all participants such that it prompted them to give an accurate albeit high-level overview of their application. The meaning of "design" was unclear to some, and this is most likely because this term has two different meanings, although both potentially relevant: a 'blueprint' of the system, or the methods used to ideate and create the system. Participants suggested expanding this requirement by adding more detailed elements such as: data provenance, how models are selected and results are evaluated, and the degree of human oversight.

Within the time limits of this project, which necessarily led to descriptions that were relatively general, most participants considered the fulfillment of this requirement as not too difficult to achieve. Some suggested that the prototype law clearly set the level of detail needed for the assessment, such as explicitly requiring 'documentation' and specifying the format of the ADIA. However, all participants rated this requirement as useful for the assessment of risks.

an assessment of the quality, integrity and representativeness of the data used to train the underlying model;

Participants were divided on the clarity of this requirement. Those who found it unclear felt that elements such as quality and representativeness were too vague as their meaning is very context dependent. One participant noted that to adequately assess the elements, it might be necessary to test the application in a real-life setting (which is at odds with the requirement to perform an ADIA before deployment). For instance, the representativeness of data can ultimately only be assessed after the application is deployed as it is *"difficult to perfectly judge the representativeness of data when the true population is not known." (NAIX)*

On the feasibility of this requirement, participants were also divided. Some participants felt that it was easy to comply with this requirement, others felt that more guidance and operationalization was needed (e.g. the aforementioned data quality and representativeness). *"Since the implementation of this particular assessment is crucial and may be subjected to different interpretations (i.e. which is the measure of data quality, integrity and representativeness), appropriate guidelines would be useful"*, Keepler pointed out. Nevertheless, most users were able to describe the methods they used to ensure quality, integrity and representativeness of the data. Among them were model-specific methods such as fairness audits, model management, explanation debugging, and generic measures for data protection such as access controls, encryption, logging. Notably, most participants – but not all – were very aware of the risks of biased data.

Overall this requirement was considered useful for the assessment of risks.

an assessment of the risks involved for natural and legal persons, with a specific focus on subjects and for end-users;

This requirement was clear for all participants on paper, although guidance on possible risks would increase the clarity even more. The assessment itself was considered difficult by participants. Most requested more guidance, for instance through examples of risks, help in performing the actual assessment, and methods to involve stakeholders in the assessment.

The most commonly identified risks were related to false positives/negatives that result from bias in training data. For example, Feedzai recognized that denying access to a bank account represents economic harm to the people being denied. As a result, their system for detecting fraud in bank account openings could disproportionately cause harm to certain groups, if not for the mitigating measures such as fairness-aware model selection and bias audits that they have implemented to maximize fairness and minimize false positives. Feedzai recognized that it is essential to carefully analyse the actual impact on people when building, deploying and maintaining fraud detection systems.

the measures envisaged to address the risks, including safeguards, security measures and mechanisms protecting the rights and freedoms of end-users and subjects, and to demonstrate compliance with this Policy Prototype, taking into account the rights and legitimate interests of those concerned.

Participants stated that the meaning of this requirement was clear to them, but many were unsure how to demonstrate compliance. More broadly, some participants were not sure about how to document the ADIA and whether they needed to share their documentation (and related assets, such as data sets) with a supervisory authority.

Most of the participants stated that they have already implemented mechanisms to ensure accuracy of decision-making. This was especially the case for participants with applications in higher risk contexts or regulated sectors such as healthcare and insurance. Mechanisms mentioned include monitoring and logging of decision-making, regular auditing of models (for instance, to detect bias), human oversight, and debugging of models by assessing local explanations.

For example, Keepler's application – which automates parts of customer support for insurance companies – is designed in a way that enables continuous auditing by humans. At any time a sample of the documents categorized by the system can be reviewed and reclassified if needed. In contrast with other automation solutions, the goal is to enhance, and not substitute, human agency.

Another example is Feedzai's 'auto-model monitoring system' for models in production that allows for early identification of missing values in specific features or an abnormal number of predicted positive instances in their fraud detection system.

Not all risk mitigation measures are technical. NAIX is an example of a participant that, alongside the technical measures that they implement, also educates their clients on both the limitations and benefits of their AI application, namely the possibility to redact vast amounts of documents that would never be possible with manual work. This enables NAIX's clients to perform adequate risk management practices while successfully using its automated application.

In week 2 of the program, companies were asked to perform and document the ADIA (Art. 4.4). The following are illustrations from templates that were provided to the participating companies for the purpose of simulating the assessment.

Requirement 1

A detailed description of the automated decision-making system, its design, its training, its data, and its purpose;

RIATLAS HEALTHCARE

Give an overview of your application and the purpose (goal) for which it is used. Describe its overall design, the data being used and, in the case of machine learning, how the model is trained.

The AI-based tool supports clinical decision-making in the patient's health status classification. Additionally, it provides smart data visualization and early detection of clinical risks to healthcare professionals in the Health Continuum Care Pathways.
The scenario is the remote monitoring (symptoms and vital signs) of patients that, after a post-acute phase, with frequent returns to the hospital (e.g. chemotherapy cycles), start a care pathway, favoring the recovery in their living environment (de-hospitalization).
Starting with data collected from heterogeneous sources, the tool classifies patient health status, using ICF (International Classification of Functioning, Disability, and Health) by WHO. At each clinical assessment of the patient, the tool identifies and suggests to the physicians (that validates them) the "appropriate" ICF codes (as a digital biomarker), in terms of functioning, activity, and participation. It also supports qualifiers (gravity level) valorization (score) for each ICF code, using recognized taxonomies and clinical assessment scales. A user-friendly visualization (intelligent dashboard) supports the monitoring of ICF codes, observing the evolution of qualifiers during the care pathway.

The tool's output is the patient health status classification, described with ICF (International Classification of Functioning, Disability, and Health) taxonomy, promoted by WHO. The output is provided as a suggestion to the physician that can accept, discard, or revise. The system incrementally trains itself by considering the feedback.

The design of the model consists of the following steps:

- **First step: we have collected clinical data from Electronic Health Record, validated from subject matter expert, and structured in a specific training dataset;**
- **Second step: we trained several machine learning models, fine-tuning their parameters for evaluating performance in terms of accuracy;**
- **Third step: we selected the most suitable model revealing good accuracy performance and satisfying clinical decision support system requirements;**
- **Fourth step: we deployed the model in the final solution, and we performed a clinical trial with patients recruited by the hospital.**

Requirement 2

An assessment of the quality, integrity and representativeness of the data used to train the underlying model;

IRIDA LABS

Describe what data is used in the application (training data and subsequent input data) and make an assessment of the data quality, its integrity, and the representativeness of the dataset.

The training is based on well-known, open-source frameworks, such as PyTorch, TensorFlow, Caffee 2. We should make clear that data, in Irida Labs' case, is video and images. The training data can be acquired from a number of sources, such as own data of Irida Labs, data collected in the field, client's data, academic datasets, and open data.

The quality of data depends on the resolution and configuration of the imaging devices. Blurred, poorly lit, or low-res data are rejected. Furthermore, data are cropped and split to pieces, in order to focus only on the points of interest. For example, if the original data is a video stream from a public road, and the desired outcome is vehicle tracking and counting, then all the clutter from the video is removed (i.e. pedestrians, background, sidewalks) and only a portion of the video is maintained; the one focusing only on vehicles. This technique reassures a high level of quality.

As far as representativeness of the data, this is a crucial factor for Irida Labs. The goal here is to make sure that the training data cover all the aspects of each particular problem, both operational as well as environmental. Moving from a baseline to an optimal AI performance requires that this problem is addressed per-case. For example, in a warehouse case where the aim is company-specific product detection, recognition and counting, all items of stock are needed for training, in multiple poses and distances, with multiple representative possibilities of other objects (i.e. people, machinery) intervening in the scene. Taking the example of a parking-lot solution (car counting, occupancy detection), all possible environmental conditions (i.e. sunshine, rain, snow) and all lighting conditions should be considered. Irida Labs' data engine has been designed around the principle that data campaigns (that is, the process of collecting the case-specific learning data) need to be as small and fast as possible, while retaining their representational quality. The automation and seamless application of this process are a key factor for real-world solutions.

Requirement 3

An assessment of the risks involved for natural and legal persons, with a specific focus on subjects and for end-users;

feedzai

Describe the risks the application may pose for subjects and end-users (see recitals 10 through 12 of the prototype for examples). Please also list risks of your application that have already been mitigated/addressed.

The fraud detection system might result in economic harm to the wrongly denied applicants, as well as unfair differential access to services. Access to banking services is paramount today, especially during a pandemic in which there has been a rapid transition to digital payments.

The study of fairness in an account opening setting is particularly important, as access to credit and other mainstream financial services that accompany a bank account often dictate a person's social mobility. It is known (and foreseeable) that underbanked communities, with difficult access to credit, have a harder time building wealth. Therefore, there's a risk that the fraud detection system will deny access to financial services disproportionately across people from different groups, based on age, place of residence, profession, or employment status.

Another risk has to do with privacy. The application contains sensitive information such as ID information, place of residence, demographic data, job and income information. All the data that we, the "developer", have access to build the ML model has been anonymized by the financial institution, the "user."

Requirement 4

The measures envisaged to address the risks including safeguards, security measures and mechanisms protecting the rights and freedoms of end-users and subjects, and to demonstrate compliance with this regulation, taking into account the rights and legitimate interests of those concerned.

keepler
DATA DRIVEN PARTNER

Describe how you would address the risks of your application through technical and/or organisational measures. Examples are testing and evaluation, monitoring of deployed models, explainability of decisions, etc. So far, you have already addressed the risks of your application, please describe what measures you would take.

To secure the data, the following measures were taken:

- **Data encryption (in transit/at rest);**
- **Logs;**
- **Data location;**
- **MFA for cloud environment login;**
- **Role-based access;**
- **Cloud auditing (log, monitor, and retention of account activity related to actions across the infrastructure).**

To ensure data integrity and to monitor the health of the deployed models, critical metrics for training, hosting and predictions are defined and collected through logs.
Each version of each model can be linked to the data used to train the models.
The system was designed to be continuously audited by humans, who are able to review a sample of categorized documents to approve or not the prediction made by the models (aKa Active Learning). This features allows for greater transparency, while ensuring a minimum level of performance through training iterations of the deployed models.

The playbook

The prototype guidance, or playbook, was introduced to participants in week 4 of the program.[32]

The playbook consists of:

- **A step-by-step risk assessment methodology;**
- **A list of values relevant to AI/ML and ADM;**
- **A taxonomy of harms;**
- **Examples of mitigating measures.**

The playbook provided participants with additional guidance both on interpreting specific concepts of the ADIA prototype law, and in suggesting a step-by-step process for conducting the ADIA framework. According to the feedback, the playbook helped participants translate the prototype law into their own contexts and made implementation more straightforward.

"It's adding practice to the theory – while ADIA details our obligations, the playbook details how to do it and what to look for." (Evo)

"The examples provided with the playbook are a very valuable resource, in some cases they have been a discussion-starter for our team that touched upon issues that have not been thought of in the past." (Irida Labs)

"Just the fact we have access to a list of values makes the ADIA more easy to conduct." (Feedzai)

In general, participants felt that the playbook clarified the ADIA by providing the much requested implementation guidance. Based on the risk assessment methodology provided in the playbook, participants identified steps that they did not take while simulating the ADIA based solely on the prototype law text.

Table 3 shows that only steps 1 and 4 were performed by all participants before receiving the playbook, which demonstrates how the risk assessment methodology from the playbook guidance provided valuable detail to the ADIA process that was otherwise overlooked. With the benefit of the playbook, for example, companies that had failed previously to identify value tensions or assess the consequences of their mitigations now knew to do so, while companies like Evoand Allegro.ai that had already done so were able to focus even more deeply on those tasks with additional guidance.

"We had not thought of determining value tensions." (Irida Labs)

32. The full prototype guidance or playbook can be found in annex 1.

Table 3
Risk assessment methodology steps taken by participants without the playbook guidance

Step	Performed
Step 1: Describe the proposed ADM system	100%
Step 2: Assess how ADM changes the existing situation	80%
Step 3: Analyse the root cause of the change	60%
Step 4: Determine impact on stakeholders and associated values	100%
Step 5: Determine value tensions	60%
Step 6: Determine probability of negative impact occurring	60%
Step 7: Identify possible changes and mitigating measures	60%
Step 8: Assess consequences of changes and mitigating measures	40%
Step 9: Decide which changes and mitigating measures to implement	60%
Step 10: Implement and document	80%

The additional elements of the playbook (list of values, taxonomy of harms, mitigating measures) were considered extremely useful. The overview of values and taxonomy of harms, in particular, were deemed to provide the guidance needed for the risk assessment. Some participants felt they were able to identify new risks of their application with the new guidance. For instance, the guidance on the value of personal autonomy helped EVO to identify possible risks to personal autonomy of their algorithmic pricing solutions which they had not identified before.

"It made me think of further potential risks, e.g. to personal autonomy, as our autonomous supply chain solution helps companies place the right product at the right time for the right price; how does this affect the person's autonomy of choice?" (Evo)

The playbook helped RiAtlas increase their understanding of the impact on stakeholders of their patient health monitoring system, determine the probability of negative impact occurring, and identify possible value tensions. Rogervoice discovered an additional utility of the playbook by using it to increase internal buy-in and understanding around the need to take the time to understand the risks of their application:

"The detailed explanation of potential risks was eye opening as well and made me think of additional ways we should mitigate risk." (Rogervoice)

All participants said they would change their ADIA after reading the playbook. This shows the need for additional guidance through operationalization and examples for a common understanding and implementation of the prototype law. Preferably, this guidance could be provided through self and co-regulatory instruments, ensuring a quick and correct uptake of the legislation's requirements.

Conclusions on policy understanding

In general, participants felt that the prototype law was clear to them on paper. However, the understanding of some elements of the prototype law varied widely between participants.

When asked about the level of understanding of the prototype law and its content, participants stated that the content was clear to them. But when asked how confident they were that they could implement the requirements listed in the prototype, only half were confident that they were able to do so. The widely shared and recurring demand from the participants for examples, operationalization of concepts, and practical guidance on the level of detail required, further demonstrated the wide gap between generally understanding the ADIA process versus practically implementing it.

The participants' appreciation of the playbook, and the fact that they would all revise their ADIAs based on its guidance, clearly demonstrated the need for specific practical guidance to complement the prototype law – and any actual future law. Such guidance can help to overcome the inherent ambiguity of norm-based regulation (which is needed for the policy to be technologically neutral), and assist in the identification and quantification of previously unrecognized risks of an ADM system.

"A chart with a scale of values can help quantify the risks." (Rogervoice)

"More detailed examples for each category would make the ADIA clearer." (NAIX)

"Concrete examples of risk quantification would help categorize low, medium, high probability vs severity levels." (Feedzai)

Assessment of policy effectiveness

To determine the effectiveness of our prototype law (our policy impact), we must determine to what extent following the requirements of the prototype law contributed to reaching our desired policy outcome. As discussed previously, in order to reach our desired policy outcome, the following four requirements would need to be met:

- **Users are able to identify what risks their application may entail for the rights and freedoms of subjects;**
- **Users are able to determine how significant these risks are (e.g. high or low);**
- **Users are able to formulate mitigating measures to these risks;**
- **Users are able to adequately assess whether these measures remove the risks or reduce them to an acceptable level (residual risk).**

Therefore, the key success indicator for our prototype was whether the prototype law had contributed to participants' identification of the risks of automated decision-making for the rights and freedoms of natural and legal persons, and also contributed to their consideration of measures to reduce those risks.

Were users able to identify what risks their applications may entail for the rights and freedoms of subjects?

All participants were able to identify risks of their application based on the prototype law. However, there were clear differences in the width and depth of the assessment. Risks related to the functioning of the application (functional risks) were identified by most participants. These are risks that relate to the performance of the model, data bias, and end-user competence. In some cases, participants had already identified these risks in the course of performing other risk assessments (for instance, a DPIA for GDPR compliance). Risks related to broader ethical or societal impacts of the application (structural risks) were identified much less frequently. Feedback on the guidance on values and harms provided through the playbook shows that it can be difficult for participants to understand and apply abstract values (for instance human autonomy) when evaluating their systems. Broadly reckoning with all the possible risks from a particular ADM application is much easier said than done, and requires strong guidance, clear examples, and ongoing practice.

The prototype aimed to make developers and users of automated decision-making systems aware of the risks their applications may pose, and enable them to find ways of mitigating these potential risks. Based on the data gathered, we can conclude that the prototype law in isolation was only partly successful in reaching its intended goal. If participants are aware of a risk, they are able to identify mitigating measures. However, identification of risks that arise from aspects not central to the purpose and functioning of the application are harder to identify and assess.

The difficulty of identifying such risks was demonstrated by the fact that many risks were not identified by participants until after receiving the guidance in the playbook, especially the overview of values and harms, and the fact that all participants said that they would revise their ADIA after seeing the playbook. For instance, by reflecting on the value of equality, RiAtlas identified a possible risk stemming from their application if insurance companies used outcomes of their patient health monitoring system to personalize insurance policies. They also determined that their patient health classification system poses a potential risk to equitable accessibility to healthcare, since it is used (among other things) to determine what care a patient needs. Rogervoice identified a risk for discrimination regarding their voice-to-text application for deaf people. This risk stems from the limited availability of voice data for certain age groups or accents, resulting in lower quality output from the system. Irida Labs identified an additional potential risk to material well-being that was not in their initial risk assessment. And guidance on the value of personal autonomy helped Evo identify possible risks to personal autonomy of their algorithmic pricing solutions which they had not identified before.

More guidance on how to assess these risks helps cast a wider net on the range of potential risks posed by AI systems, while providing better tools to identify and assess the most relevant ones. The list of values and the taxonomy of harms, in particular, helps ensure that a broad set of possible risks are being considered, and the step-by-step risk assessment process ensures that the most significant ones are effectively flagged and addressed.

That said, it was sometimes unclear for participants which values, rights and freedoms could potentially be affected by their application. The overview of values and harms in the playbook enabled participants to think about how these concepts related to the use and broader societal effects of their own applications. This complemented the more technical and application-centric

perspective to risk assessment that most participants brought to the table, according to which the risk tended to be associated with the functioning and operation of the AI system. In other words, the playbook and its resources reversed the participants' default risk assessment strategy, shifting their thinking process from focusing only on what technical and functional risks their application may pose, to reflecting on how structural and societal risks might emerge and manifest themselves through the deployment and use of the application. The playbook and its various elements enabled participants to reflect on a potential wider set of risks, conceptualized through lists of values and types of harms, and on how these risks could occur through the use and interaction of their applications in society. In this way, participants were no longer exclusively constrained by the technical underpinnings of their applications to derive and identify risks, which helped them identify unknown risks related to a broader set of structural and societal effects of their applications.

"[The list of values and harms] are useful because they allow us to analyse the problem in an exhaustive way and from different points of view." (RiAtlas)

An insightful comment on the prototype law from one of the participants was that the ADIA did not require any justifications of the choices made. Rather than just describing and documenting their findings, they suggested that a user of an ADM system should also describe why particular risk-reducing measures were taken (and others not), and how these measures reduce the risk to an acceptable level, or take it away altogether.

Another insight was that it may be helpful for the playbook's guidance to be more directly informed by other fields' approaches to risk assessment. For example, the AI field might be able to derive useful lessons from how environmental impact assessments are used in other industries to evaluate the impact of certain chemicals or industrial processes.

Were users able to determine how significant the identified risks are (e.g. high or low)?

Given that the prototype law did not have an explicit requirement to determine the significance of the risks identified, we did not gather sufficient information regarding the ability of users to proceed and perform such determination. The prototype law establishes a requirement to assess the risks, and a requirement for users to consult with the supervisory authority in cases where the impact assessment indicates that the application may result in a high risk and these risks cannot be mitigated. The intermediate step of assessing the significance of the risk identified is implicit and, as demonstrated through the program, was not performed by most of the participants. Whether this is due to a lack of awareness or a lack of understanding on how to perform this assessment was unclear. Either way, this finding demonstrates the need for a more descriptive and explicit procedure to determine the significance of the risks posed by AI systems and applications.

Were users able to formulate mitigating measures?

Participants were confident they were able to identify appropriate risk reducing measures. Those that have performed DPIAs for their systems (which also require risk identification and mitigation) were particularly confident and felt that not much additional work was needed. For instance, NAIX already has a process in place for assessing risk, implementing risk reducing measures and determining residual risk, into which they incorporate our prototype requirements. This was made easier by the fact that our ADIA prototype to some extent mimics

the DPIA requirement from article 35 of GDPR. Most of the risk-reducing measures identified by the participants were focused on ensuring the accuracy and fairness of the automated decision-making and the protection of personal data. Feedzai, for instance, proposed the following concrete risk reducing measures:

- ***"Fairness-Aware Model Selection: there might be a large spread over the fairness metric at any level of predictive accuracy, and therefore we select the model that goes to production based on the optimal fairness-accuracy tradeoff.***
- ***Auto-Model Monitoring: we have proprietary algorithms for continuous monitoring of models in production that allow early identification of missing values in specific features or abnormal number of predicted positive instances.***
- ***Frequent Bias Audits: bias can creep in anytime so we establish frequent (monthly) bias audits to assess if there is any fairness degradation in our model, and if there's a need to retrain.***
- ***Debugging through Explanations: we ask both data scientists and fraud analysts to debug a sample of predictions of the ML model using post hoc explanation methods. We monitor if feature attribution changes over time."***

Were users able to adequately assess whether these measures remove the risks or reduce them to an acceptable level (residual risk)?

While the participants felt they could provide risk-reducing measures for their application, they were unsure how to assess the effectiveness of those measures. For example, RiAtlas was very confident that they could apply the ADIA principles and identify risks, but was less confident in evaluating the adequacy and fairness of measures to reduce the risk. Feedzai noted the related concern that the lack of explicit guidance around how to assess the effectiveness of mitigation measures may leave too much discretion in the hands of users. This concern is exacerbated when considering the complex question of how to identify and assess mitigations that address broader ethical impacts or societal risks from the application, as opposed to narrower, functional risks that are more directly related to the technical operation of the application.
Another example demonstrating the difficulty in evaluating the effectiveness of mitigating measures came from Reface, which identified the risk of misuse of their face swapping technologies for the creation of misleading content. The measures that could help mitigate this risk included applying watermarks, reviewing content generated by end-users and setting community guidelines. However, assessing the effectiveness of these measures after deployment of the application would be very difficult, while assessing them before deployment (which the ADIA requires) would be even harder.

Some participants' challenges with assessing the effectiveness of mitigating measures were exacerbated by the fact that the balancing of values and interests, and the reduction of risk to an acceptable (residual) level based on that balancing, were not specifically mentioned as mandatory elements of the risk assessment as defined in article 4.4 of the prototype law. This led some to conclude without sufficient analysis or documentation that they had properly managed their risk. This gap could be addressed by adding a specific requirement to document that balancing and justify the residual risk in the ADIA.

Assessment of policy costs

The costs of implementation of the ADIA policy prototype can be understood as a combination of the time and resources required to perform the ADIA, the costs of implementing mitigating measures, and other compliance efforts (e.g. monitoring). These are the direct costs associated with complying with the requirements set by the policy.

There may also be indirect costs. For instance, changes to the application may impact revenue in a negative way. However, the impact on revenue may also be positive if trust in the application grows.

Given the four-week duration of the policy prototyping exercise, we focused our questions on the immediate direct costs associated with the performance of the ADIA (i.e. how much time and resources did it take to perform the ADIA). When we asked participants about their estimate for performing an ADIA for their application, the most prevalent range was between 5-50 hours.

Regarding the roles involved by participants to perform the ADIA, it is clear that the ADIA requires an interdisciplinary team. All participants would involve both legal and technical functions, with some including risk and compliance functions as well. Besides internal functions, some participants would require outside counsel to be able to perform the assessment.[xvii] The involvement of such external experts can be particularly costly for smaller organisations such as startups.

We may conclude that costs of implementing the ADIA are dependent on many factors related to the type of organisation (e.g. size, maturity), the ADMs in scope of the regulation, and system characteristics. The need to involve various internal functions and, in some cases, external counsel for performing the ADIA is a major component of the implementation cost. However, while a significant investment of time or resources would clearly be necessary to conduct an ADIA, we do not have concrete evidence from the participants indicating that this requirement would be overly burdensome. Nor were there any significant unforeseen effects associated with the performance of an ADIA that would lead to additional costs for participants or other stakeholders.

Discussion and way forward

06

ADIA prototype policy evaluation

Results and observations

Based on the assessment of our three criteria (policy understanding, policy effectiveness, and policy costs), we can conclude that, overall, our prototype law has been successful in achieving the desired policy outcome:

Developers and users conduct effective risks assessments for their ADM application.

Policy Understanding

First, the prototype law was sufficiently clear in its wording for users to develop a basic understanding of much but not all of the tasks they were required to do, namely:

- identify what risks their applications may entail for the rights and freedoms of subjects;
- determine how significant these risks are (e.g. high or low);
- formulate mitigating measures to these risks;
- assess whether these measures remove the risks or reduce them to an acceptable level (residual risk).

One of the sources of misunderstanding was the prototype law's description of the types of actors involved. Participants did not always identify themselves as a developer or a user, for instance. Furthermore, the complex landscape of AI actors means there may also be dependencies between actors. For instance, the limitations of pre-trained models and the specificities of AI/ADM platforms known to developers are also important to know for users 'downstream' who would conduct an ADIA. Also, the complexity of the AI landscape makes it less clear who is responsible and capable of executing an ADIA. For instance, a developer of learning algorithms might not have knowledge of or control over training data selected by the user or the purposes for which the model is used. This makes it hard (if not impossible) for them to conduct an ADIA.

Policy Effectiveness

Based on the outcomes of the ADIA exercise and the feedback of the participants, we can also conclude that, overall, our prototype law has been effective. In general, the most important requirements for the desired policy outcome were met by the participants – that is, they were able to identify risks posed by their applications (requirement one) and formulate mitigations to address those risks (requirement three). However, most participants did not understand that, as a part of their assessment, they were also required to determine which risks were high or low in order to help inform their mitigation decisions (requirement two), and to assess their mitigation measures' effectiveness at reducing high risks to an acceptable level (requirement four). In both these cases, the addition of explicit requirements in the prototype law and more concrete guidance in the playbook was clearly necessary to foster greater policy understanding and effectiveness.

Policy Cost

As with most regulatory requirements, there would certainly be non-trivial costs involved in fully complying with an ADIA requirement. And although our prototyping program was necessarily limited – as noted previously, the ADIA process tested with participants was shorter and less detailed than what would likely be required for compliance with an actual law – we did not get the impression from the participants that conducting an ADIA would overburden them. One key reason for this was the overlap between the ADIA process and the GDPR DPIA requirement that many of the companies already comply with, such that the results from one can be reused for the other. This suggests that policymakers should focus on deliberately integrating ADIA and DPIA requirements in law to avoid unnecessary or duplicative costs for developers and users.

Ability of the Prototype to meet the desired policy goal

As expressed in the anonymized survey at the end of the program*, participants felt that the ADIA process gave them new insights into the risks of AI / ADM applications:

"The ADIA process helped us to think about new insights into potential risks, in terms of privacy & data protection, fairness and rule of law."

"We realised that we should incorporate more controls in our current workflow to evaluate all the potential risks of a new application."

"The adoption of the ADIA process in our organisation helped to understand the implication of the AI developed tool in the real world. It also raised awareness on possible risks and this helped to do that in a more responsible manner."

As discussed in the same survey, several participants are contemplating using the prototype law and the associated guidance to help improve their existing risk assessment policies and processes:

"The taxonomy of harms is particularly useful and we are considering adopting it in our internal processes."

"We are revising our current risk assessment processes to include some of the features learned during the ADIA."

"We are thinking about creating a more standardized ADIA instead of doing it based on case by case requirements depending on geography."

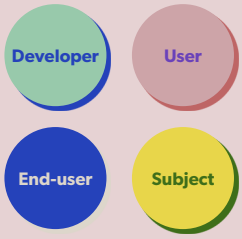
In the final co-creation workshop, participants also made clear that they would consider disclosing ADIA documentation to demonstrate trustworthiness to their clients and to differentiate from competitors.

From this range of feedback, we may further conclude that the prototype law has contributed to the policy goal.

*Quotes on this page stem from the final evaluation survey of the program. The survey aimed at obtaining an overall reflection of the program experience and identifying both operational and strategic implications that participating companies foresaw for their business from the ADIA prototyping journey in our program. This survey was anonymized.

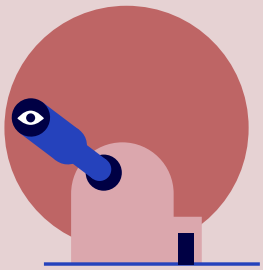
Possible improvements to the ADIA prototype law

While the policy seemed to be effective overall, based on the results of the exercise and the feedback of the participants, there is definitely room for improvement. The following elements could be considered in order to further improve the prototype:



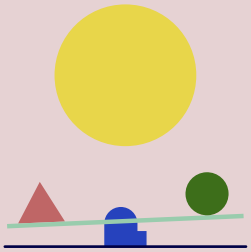
Description of actors

The AI landscape is more complex than anticipated in the prototype. The prototype law's distinction between developers, users, end-users and subjects is helpful, but does not capture the full complexity of the AI landscape. Furthermore, the terminology used in the prototype differs from that generally used in the tech community. For instance, a 'user' in the tech community is generally the person using the service, not the organisation using/deploying the ADM system.



Greater focus on justification and legitimacy

The requirements that an ADIA should meet (article 4.4) did not specifically require participants to justify the use of their application and the adequacy of their risk-reducing measures. By including these as clear requirements, a more conscious balancing of interests – and a documentation of that balancing – could likely be achieved. For instance, developers and users could not only document that they are taking measures to reduce bias in datasets, but also how those measures are implemented and why they are sufficient to address the risks. This type of justification could help assess the effectiveness of the mitigating measures, which was reported by the participants as one of the main difficulties they experienced in the process.



Greater emphasis on value tensions as part of the ADIA requirements

Given how the playbook's guidance on discovering possible tensions between the values posed by AI systems helped participants to identify additional risks, this particular step should be added to article 4.4 as a clear requirement of the ADIA prototype law. This is closely connected with the previous recommendation that the justifications for the selection and adequacy of risk mitigation measures should be documented.

Specific changes to the ADIA prototype law

Based on the results and the feedback, the prototype can be revised and thus improved in detail. For article 4.4 (the most important article in the prototype), the following changes could be made:

	Amended text based on feedback	Comments
4.1	**Prior to the deployment of an automated decision-making system, the user shall assess the risks of the envisaged automated decision-making system and its application on the rights and freedoms of natural and legal persons.**	**The definition of an automated decision-making system must be revisited in the definitions (article 3).**
4.2	**In those cases where the application of an automated decision-making system is likely to result in a high risk to rights and freedoms of natural or legal persons, the user shall carry out an automated decision impact assessment prior to the deployment.**	
4.3	**An automated decision-making impact assessment is mandatory in all cases where there is high risk. Examples of high risk, include, but are not limited to the following situations: ~~an automated decision impact assessment referred to in paragraph 2 shall in any case be required in case of:~~** • **potential ~~unfair bias~~ or discrimination towards subjects, including price discrimination, employment discrimination or unfair differential access to services;** • **potential loss of control or agency for the subject, including economic or psychological manipulation;** • **large scale application of automated decision-making, including profiling and systematic monitoring, that may affect communities or society as a whole.**	**Rephrased to avoid confusion that this is a limitative list. It could be argued that this paragraph could be left out entirely if more guidance is provided alongside the prototype.** **Removed the term "unfair bias" given its ambiguous meaning. It was not clear whether the text was referring to "bias" in a statistical sense, to model or label bias in an ML fairness sense; or plain language "bias."**

Continues next page...

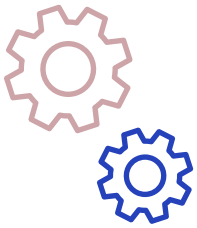
	Amended text based on feedback	Comments
4.4	An automated decision-making system impact assessment shall contain at least: • a detailed description of the automated decision-making system, its design, its training, its data, and its purpose; • an assessment of the quality, integrity and representativeness of the data used to train the underlying model; • an assessment of the risks ~~involved~~ the automated decision-making system poses to the rights and freedoms of ~~for~~ natural and legal persons, with a specific focus on subjects and for end-users; and a determination of how significant the risks are. • A description of the measures envisaged to address the risks including safeguards, security measures and mechanisms protecting the rights and freedoms of end-users and subjects and to demonstrate compliance with this Regulation, taking into account the rights and legitimate interests of those concerned, and an explanation why these measures are deemed adequate. • An assessment of the legitimacy and necessity of the deployment of the automated decision-making system.	Added an explicit reference to the need to determine how significant the identified risks are. More focus in the requirements under 4.4 on the value tensions and the critical reflection on the adequacy of risk reducing measures and legitimacy via an additional requirement (justification).
4.5	In those cases where the automated decision impact assessment indicates that the application may result in a high risk to the natural rights and freedoms of natural and legal persons and these risks can or will not be mitigated, the user shall, prior to the deployment, consult with the supervisory authority.	

Recommendations

07

ADIA prototype policy evaluation

Recommend-ations for regulating AI/ automated decision-making



Based on the results of the prototyping exercise, and the feedback on the prototype law and playbook, we would advise lawmakers dealing with the question of how to develop a risk-based approach to AI regulation to take the following recommendations into account:

Focus on procedure instead of prescription as a way to determine high risk AI applications

- In the context of risk assessment, a prescriptive approach classifies *a priori* the set of risks that organisations are required to identify. It does so by stipulating a rigid *ex-ante* list of high-risk applications defined based on a given criteria such as sector, intended use, etc. The prescriptive approach is, in a way, automatic: if a given AI application falls in the list, it will be considered high risk. A procedural approach enables organisations to identify, assess and mitigate risks by following a number of steps, indicative criteria and examples. The procedural approach is not automatic and it is not directed solely at the identification of risks. It involves the carrying out of a methodical process through which the risks will not only be identified, but also mitigated. In effect, organisations should rely on their own corporate ethics and values; internal governance structures and measures; internal roles, teams, and responsibilities; operations management; and strategies for communicating with external stakeholders to determine the risks posed by AI systems. The ADIA framework is an example of a procedural approach.
- The findings of the program confirm the importance and usefulness of codifying a risk assessment procedure as a viable governance mechanism. Putting in place a procedure to identify, assess and mitigate risks, accompanied by detailed guidance (taxonomy of values, examples of harms and list of possible mitigation measures), enables organisations to better understand, document and address the risks posed by their AI systems. As demonstrated by this program, and while acknowledging the limitations of this exercise, the participants – startups operating in different regions and sectors – were able to adopt and implement the ADIA framework.
- Based on the results of the program, a process-based risk assessment approach to determine high risk AI applications seems to be a sound and workable alternative to a rigid and prescriptive approach based on a combination of sectors and intended uses.[XVIII] A step-by-step risk assessment approach – unconstrained by prior sectoral determinations and complemented by a set of examples of risks and taxonomy of values – will do a better job at helping

organisations assess risks based on the specific context and impact of proposed AI uses.[XIX] This procedural approach will also do a better job at taking into account the dynamic and iterative character of AI, where systems are continuously evolving and changing as a result of their interactions with people and the environment. As risks posed by AI systems may change as they keep evolving, the encapsulation of risks based on generic sectoral assumptions is an ill-defined solution. Unlike a procedural methodology, a prescriptive approach will struggle to identify and regulate emerging risks in such dynamic conditions.

- One should also take into account that there is a higher level of uncertainty and complexity in ascertaining certain types of risks posed by AI/ML systems than others. As demonstrated through this program, broader structural or societal risks grounded in moral or ethical values (in contrast with functional risks based on the operation of AI systems) are difficult to identify, assess, and mitigate. This additional complexity requires robust step-by-step procedural approaches to risk assessment, complemented with operational guidance, rather than an approach anchored on rigid classifications based on the sector in which the AI is being utilized.

- Notably, the European Commission is proposing a procedural approach to risk assessment and management in the context of its recent Digital Services Act (DSA) proposal.[33] The latter lays down obligations for very large online platforms to conduct risk assessments on the systemic risks brought about by or relating to the functioning and use of their services, and to take reasonable and effective measures aimed at mitigating those risks. The DSA act also foresees additional transparency reporting obligations, which include a report setting out the results of the risk assessment and the related risk mitigation measures identified and implemented. Both in the context of its AI proposal and the DSA proposal, we would urge the Commission to consider the results of this policy prototyping experiment, and to align on a consistent procedural based approach to risk assessment throughout its various regulatory proposals, avoiding in this way an inflexible prescriptive approach like that proposed in its AI White Paper.

Leverage a procedural risk assessment approach to determine what is the right set of regulatory requirements to apply to organisations deploying AI applications (instead of applying all of them by default)

- A procedural approach also has the virtue of acknowledging and relying on factors related to the nature, severity, probability, and reversibility of potential harms; the opportunity for individuals to exert control over or opt out; and the extent of human oversight and level of automation of a given application. This enables a more granular assessment of the degree of risks posed by AI,[34] and consequently enables a more granular determination of what corresponding mitigation measures are necessary and appropriate. If one equates such mitigation measures to the set of regulatory requirements companies should follow when building and deploying their applications, the procedural angle enables a more balanced and adaptable regulatory approach. Rather than applying an entire set of regulatory requirements by default –

33. EC 2020c.
34. Such level of granularity is associated with the fact that the procedural approach is also more attuned to incorporate qualitative insights in its risk assessment (see section above on qualitative type of risk assessments).

and regardless of the type of AI application, its context and actual risks – the procedural approach allows for a more flexible and appropriate application of regulatory requirements like human oversight, explainability, rights of redress, monitoring, and disclosure requirements, amongst others. Through such an approach, statutory requirements would not be assigned in bulk based on an inflexible list of sectors or applications, as proposed in the European Commission's AI White Paper, but instead would be tailored to the specific AI application in question and the level and extent of the risks assessed, weighed alongside a calculation of the benefits the application brings.[XX]

- The procedural approach acknowledges the importance of looking at the specific context in which an AI application is being built and planned to be deployed, and helps determine and tailor the application of specific regulatory requirements to high risk AI applications based on that context.[XXI]

Provide specific and detailed guidance on how to implement an ADIA process, and release it alongside the law

- The examples of AI risks helping companies understand whether an ADIA should be considered, both in article 4.3 and in the playbook, proved to be useful. This showcases the need for further guidance, namely in terms of lists of examples and assumptions, to ensure a consistent and reliable risk assessment process.

- Given how participants received, appreciated, and used the playbook, we strongly encourage the provision of additional guidance on how to interpret and implement any ADIA requirements – or any other regulatory requirements for high-risk AI, for that matter. Ideally such guidance could be provided through soft law or co-regulatory instruments, in order to ensure appropriate flexibility and adaptability to changes in technology and society. The playbook's step-by-step process and its additional elements (list of values, taxonomy of harms, mitigating measures) were considered extremely useful by the participants. The overview of values and taxonomy of harms, in particular, provided suitable and needed guidance to conduct a proper AI risk assessment. This demonstrates how additional operational guidance accompanied by examples helps foster a common understanding, interpretation, and implementation of the law. This also demonstrates that policy guidance is particularly appreciated when it is "accessible", i.e. made and phrased in a manner that allows for it to be understood and used by those who are actually developing and deploying AI systems in practice. [XXII]

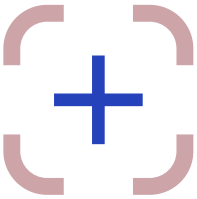
- While the recitals provided some context for the prototype law, the overwhelming majority of the participants stated that they benefited from the playbook and would have actually changed their risk assessment had they had access to the playbook. The ADIA playbook not only gave participants new insights into the risks of AI/ADM applications, it also enabled some of them to identify new risks. Thus, we conclude that guidance on how to comply with any new requirements should be provided simultaneously with any new legislation defining those requirements, rather than

provided *ex post* through interpretation of requirements by supervisory authorities or the courts, to provide more clarity and certainty to norm addressees. This guidance could, for instance, be framed in the form of guidelines, an ADIA template, or a step-by-step compliance guide.

- Additional guidance can also provide the necessary tools to tackle both narrow functional or technical risks and broader structural or societal risks. The overview of values and harms in the playbook enabled participants to think about how these concepts relate to their own applications. This guidance complements the more technical and application-centric perspective to risk assessment, according to which the risk tends to be associated with the functioning and operation of the AI system, and helps identify unknown risks. In other words, the playbook and its resources reverse the risk assessment strategy, shifting the thinking process from what risks an application may cause through its technical (mal)function or (faulty) operation, to how risks could emerge and manifest themselves through the deployment, use and interaction of an application in society. This shift enables AI developers to reflect on how risks may affect broader societal values. Obviously, not all risks will emerge and manifest themselves through the use of all applications, but the combination of these two approaches – technical and value driven – helps to identify new relevant risks. Some participants also felt that this approach of assessing their application based on a list of values made the process more 'objective and systematic'.

- Guidance, such as that provided through the ADIA playbook, helps overcome the inherent ambiguity of norm-based regulation (which is needed for the policy to be technologically neutral) and, through taxonomies and examples, helps identify previously unknown aspects of an ADM system.

- The demand for additional guidance from participants' also confirmed the need for a tighter calibration and coordination between different governance instruments – hard law, soft law, and co-regulation – to ensure a regulatory regime and guidance that is comprehensive while still being flexible, adaptable, and deeply informed by the practical experience of AI practitioners and other relevant stakeholders.



Be as specific as possible in the definition of risks within regulatory scope

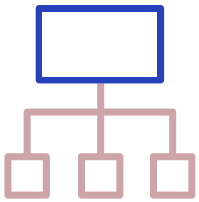
- As noted throughout the report, assessments of risk need not focus solely on the technical or functional concerns associated with AI/ADM (e.g. explainability, transparency, accuracy of decision-making), but can also include assessment of the broader structural concerns and societal impacts that may be associated with AI/ADM (e.g. overreliance on AI/ADM systems, infringement of human rights, dehumanisation, impact in terms of fairness, proportionality, societal effect).[XXIII] This is in line with not only how the technology operates, but also on who is impacted by it and how.
- However, the feedback we received from the participating companies revealed that risks related to the functioning of AI systems (how they are built and operate) are much easier to identify than risks related

to the application and consequences produced by these systems. In particular, the feedback on the guidance on values and harms provided through the playbook shows that it can be difficult for participants to understand how their products or services may implicate abstract values (for instance, human autonomy). Such difficulty prompted one of the participants to note that *"the primary challenge here is to be imaginative enough in contemplating the types of harm that these systems could result in and assessing the risks of such harm in a meaningful way."* If applying a law requires *"moral imagination"*, then that law runs the risk of not being clear and not offering enough legal uncertainty. To avoid this level of uncertainty, we urge policy and lawmakers to work with academia, civil society, and industry to clearly specify the types of risks and harms they expect to be identified in a systematic manner and mitigated in an effective way.

- A playbook like the one used in our ADIA framework is a good first step to reduce this uncertainty and avoid the burden of trying to identify and solve every possible "moral implication" of AI-based products and services. Such additional guidance on how to assess these risks can provide a taxonomy of potential risks posed by AI systems, along with tools to better identify and assess the most relevant ones. A list of values and the taxonomy of harms, in particular, helps ensure that an explicit set of possible risks are being considered; and the step-by-step risk assessment process ensures that only the most significant ones are effectively flagged and addressed. Given the difficulties reported by the participating companies, however, the prototype playbook could be improved to provide more specific guidance on how to identify, assess, and mitigate risks related to ethical issues and societal impacts. Based on the feedback given by the participating companies, any new law or guidance around AI risk assessment will need to clearly and narrowly specify the types of risks that it is targeted at, to ensure that organisations are able to understand and practically comply with it. Yet providing such clarity in regard to broader ethical or societal risks will necessarily be challenging, as other commentators have pointed out.[XXIV]

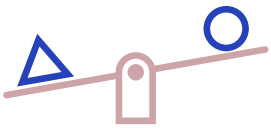
Improve documentation of risk assessment and decision-making processes by justifying the selection of mitigation measures

- Deciding and documenting how to mitigate risks posed by AI systems needs to be part of any AI risk assessment process, and is a fundamental element informing the overall AI risk-based approach.
- Based on the feedback received by our program participants, it would be helpful if users (deployers) of an ADM system also described in their ADIA why particular risk-reducing measures were taken (and others not), and how these measures reduced the risk to an acceptable level (or removed it altogether). The reasons for accepting any residual risk should also be included in the ADIA. Providing these further insights on the value and effectiveness of the risk-mitigating measures selected would help determine the right set of regulatory requirements applicable to the AI application in question, and bring greater clarity as to how tensions amongst values affected by AI/ADM are resolved.



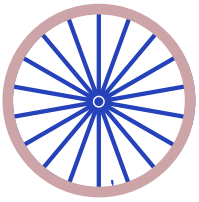
Develop a sound taxonomy of the different AI actors involved in risk assessment

- When regulating AI/ADM, lawmakers must be cognizant of the complex landscape of actors involved in developing, deploying, using, and being impacted by AI/ADM. The responsibility to conduct an ADIA may be shared by different actors. A taxonomy that reflects these different roles and clarity on who is responsible for conducting ADIAs (or which parts of an ADIA) is recommended.
- The development of such taxonomy is important for two main reasons: 1) to appropriately assign the tasks of identifying, assessing, or mitigating risks; and 2) to better understand the group of stakeholders being affected by ADM.



Specify, as much as possible, the set of values that may be impacted by AI/ADM and provide guidance on how they may be in tension with one another

- When implementing a requirement to do a risk assessment for AI/ADM, it is important to be very clear what is desired. In particular, guidance and explanation on values that may be affected by AI/ADM and value tensions that may arise are very helpful. More experience in understanding the impacts of ADM might also lead to better identification of risks.



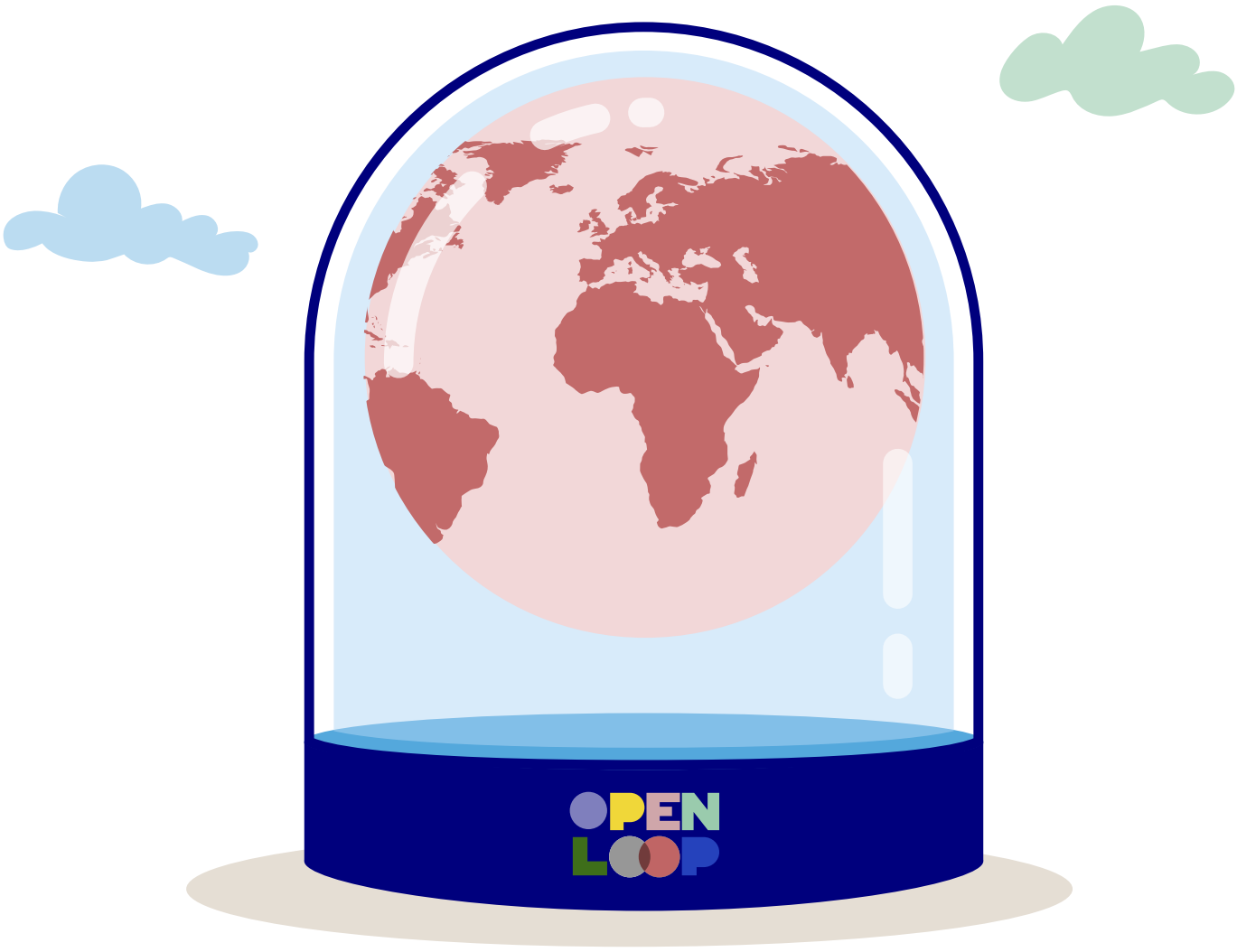
Don't reinvent the wheel; combine new risk assessment processes with established ones to improve the overall approach

- Throughout the program, participants had already identified risks that relate to the performance of their model, data bias, and end-user competence, in the course of performing other risk assessments (for instance, a DPIA for GDPR compliance).
- In many cases, there is an overlap between the ADIA and the GDPR DPIA requirements. In order to avoid double work and costs, ADIA and DPIA requirements should be integrated in law.[xxv]

Final reflections on the policy prototyping methodology

Testing our prototype law with participants has yielded some key insights into the value of introducing a risk assessment process for AI/ADM as a valid governance option within the evolving regulatory debate. In keeping with the idea of prototyping, lawmakers can take our experiences with this policy prototyping exercise and improve on it. Through our Open Loop program, we encourage lawmakers and regulators to embark on more large-scale prototyping exercises that could surpass the limitations of this current program and test with a greater degree of accuracy the effectiveness of a mandatory AI/ADM risk assessment in practice.[xxvi]

This policy prototyping program also helped us test the concept and assess the value of policy prototyping itself. As a methodological instrument aimed at producing evidence-based recommendations to policymakers, we found this program to be a promising avenue for multi-stakeholder collaboration, and an agile vehicle to co-create, test, iterate, and shape the governance of new and emerging technologies. We look forward to the possibility of similar experiments in the future with an even broader set of partners and participants.



Come join us!

End notes

I. This is the case with the European Commission (2020a), which proposed in its White Paper on AI that an AI application would be deemed high risk if it meets both of two criteria: (1) *"the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur," and (2) "the AI application in the sector in question is […] used in such a manner that significant risks are likely to arise."* (p. 17)

II. A good example of a multi-tier approach to risk determination comes from the German Data Ethics Commission's opinion on algorithmic and data governance (2019), which proposes a five tier risk classification based upon a combined "severity x likelihood" calculation, with different levels of AI regulatory obligation attaching to each different level of risk, ranging from (level 1) no additional regulatory obligations to (level 5) prohibition of the application. Another example is the one provided by the Center for Democracy and Technology (2020) in its response to the EC White Paper on AI, advocating for multi-tiered risks based on severity and likelihood.

III. This is again the case with the European Commission's AI regulatory outline presented in its White Paper (EC 2020a). For the sector criterion, the EC suggests that the "sectors covered should be specifically and exhaustively listed in the new regulatory framework. For instance, healthcare; transport; energy and parts of the public sector" and that *"[t]he list should be periodically reviewed and amended where necessary in function of relevant developments in practice."* (p. 17)

IV. A good example of the qualitative approach can be found in Singapore's AI Model Governance Framework (2020) and its Companion Guide (2020), which present a long list of questions that organisations should consider related to AI risk, with the goal of collecting stakeholder feedback and encouraging dialogue and reflection about the AI risks and mitigations. The Dutch government in collaboration with Considerati have published their proposed Artificial Intelligence Impact Assessment (2018), which similarly lists a set of questions meant to elicit risk analysis and determinations. The IEEE's Standard 7010-2020: Assessing AI Impact on Human Well-Being (2020) is another highly qualitative approach to measuring risk, in that it presents a long-term, life-cycle assessment of AI applications and calls for ongoing monitoring, revision, and iteration of both the AI and the Impact Assessment itself. The AI Now Institute has proposed an Algorithmic Impact Assessment (AIA) framework, designed for public agencies and aimed at supporting affected communities and stakeholders as they seek to assess the claims made about automated decision systems, and to determine where – or if – their use is acceptable. (add also link for the term Algorithmic Impact Assessment (AIA) referenced above: https://ainowinstitute.org/aiareport2018.pdf. And from the academic side, Calvo et al (2020) have proposed a "Human Impact Assessment for Technology" in their Advancing impact assessment for intelligent systems, which introduces social science methodologies and gathering of qualitative input from the stakeholder population into the risk assessment process.

V. This is the case of the Algorithmic Impact Assessment tool (2020) that is being developed by the Canadian Government to help federal agencies comply with the Directive on Automated Decision Making. The tool is being developed through a public-private partnership, with an open source license, and is currently hosted on Github. The tool is uniquely quantitative, consisting of approximately 60 different questions (most of them requiring only a simple yes or no answer) that impact the risk score. Half of the questions are "impact questions," the answers to which incrementally increase the total risk score of the project. The other half of the questions are "mitigation" questions which incrementally decrease the risk score of the project. (See the full list of questions here, as of November 2020).

VI. Qualitative assessments have support from many regulatory agencies and government-affiliated groups. The Assessment List for Trustworthy AI (ALTAI), tested and proposed by the EU's High Level Expert Group on Artificial Intelligence (HLEG-AI), EC 2020b, consists of a lengthy set of questions that organisations should address for any new AI application that has significant impact on human lives (*e.g.* potential to interfere with fundamental rights, or applications that present physical safety risk). The intention of the ALTAI, however, is not to produce a risk score, or even a final determination as to whether the AI is high risk. Rather, the ALTAI is meant as a collaborative reflection exercise, focusing organisations on the most important questions to answer and issues to address before deploying a new AI. The IEEE's (2020) Standard 7010-2020: Assessing AI Impact on Human Well-Being is also a highly qualitative approach to measuring risk, in that it presents a long term, life cycle assessment of AI applications and calls for ongoing monitoring, revision, and iteration of both the AI and the Impact Assessment itself.

End notes

VII. See the EU Agency for Fundamental Rights' report "Getting the Future Right: Artificial Intelligence and Fundamental Rights" (2020), recommending the EU legislator to consider making mandatory impact assessments that cover the full spectrum of fundamental rights. See also Mantelero's (2020) AI and Big Data: A blueprint for a human rights, social and ethical impact assessment, describing and proposing a broad AI impact assessment intentionally modeled off of DPIAs, but expanded to include human rights and ethics related to issues not normally encompassed by DPIAs; the Center for Democracy and Technology's Response to EC (2020), proposing that a separate human rights impact assessment (HRIA) should be conducted on top of an AIA for any application that presents risks for individual liberties or rights; the Data and Society's Governing Artificial Intelligence: Upholding Human Rights and Dignity (2020), arguing that human rights should be the central lens for thinking about the harms that could occur from AI, and the IEEE 7010-2020 – Recommended Practices for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being (2020), which incorporates many concepts from human rights law into a somewhat broader "human well-being" framework.

VIII. See Facebook's response to EC White Paper on AI (2020), which presents the concept of Automated Decision Impact Assessment (ADIA), akin to DPIAs, as a *"more balanced alternative to requiring blanket prior reviews by a regulator" (p.4) and as a way to "align with GDPR's principle of accountability whereby organisations acting as controllers are in the best position to assess, determine, and document the level of risk raised by their own processing activities."* (p.18). See also Submission of the Chair Legal and Regulatory Implications of Artificial Intelligence of Université Grenoble Alpes to the EC's White Paper on AI (2020), stating that *"[c]ompliance of high-risk application with the legal and ethical requirements should first be self-assessed by the developers themselves. We could draw here a comparison with the Data Protection Impact Assessment (DPIA) existing under the GDPR under the 'privacy by design' principle."* (p.16)

IX. Daniel Schiff et al (2020) address this question in their academic paper, Principles to Practices for Responsible AI: Closing the Gap. The paper offers five (interconnected) reasons to explain the difficulty in developing an AI Impact assessment. (1) The scope and complexity of potential impacts is so vast (environmental, democratic, physical safety, human agency, economic, etc.) that it becomes hard for any single tool or process or resource to help an organisation successfully consider, identify, and mitigate risk across the range of possible impacts. (2) So many different disciplines are involved in the creation of AI systems that organisations and regulators find it difficult to know where to place accountability for the systems. (3) Each different discipline that is involved in creating and monitoring AI have different (non-harmonious) concepts of AI risk and mitigation techniques. (4) There is an abundance of AI risk assessment tools being developed and promoted by various organisations, many of them (to date) lacking real evidence for their effectiveness, and each of them distracting from a consensus. (5) The functional separation of technical and non-technical experts within organisations limits the potential to communicate effectively, understand issues robustly, and respond to considerations of AI impact on well-being.

X. The idea of emulating AI Impact assessments on the existing DPIA model has also been proposed by a number of academics. See, *e.g.* the Submission of the Chair Legal and Regulatory Implications of Artificial Intelligence of Université Grenoble Alpes to the EC's White Paper on AI (2020): "Compliance of high-risk application with the legal and ethical requirements should first be self-assessed by the developers themselves. We could draw here a comparison with the Data Protection Impact Assessment (DPIA) existing under the GDPR under the 'privacy by design' principle."; Mantelero's (2020) AI and Big Data: A blueprint for a human rights, social and ethical impact assessment, who describes and proposes a broad AI impact assessment intentionally modeled off of DPIAs, but expanded to include human rights and ethics related issues not normally encompassed by DPIAs); and Katyal's (2019) Private Accountability in the age of Artificial Intelligence, who similarly describes an assessment model for AI – the Human Impact Statement – which is based on the DPIA model, but expanded to include population level and societal impacts that are not well addressed by DPIAs.

XI. As noted in the paper, Facebook believes that such an approach would be a more balanced alternative to requiring blanket prior reviews by a regulator of all "high-risk" AI applications as the EC White Paper recommends, and that would also align with GDPR's principle of accountability whereby organisations acting as controllers are in the best position to assess, identify, document and mitigate the risk raised by their own processing activities. Given the existing processes and operations companies have already created to conduct DPIAs, one could add that there is already precedent and familiarity with this type of self-assessment that could be adapted to AI regulation.

End notes

XII. As noted in Facebook's response, "[c]onsistent with GDPR's approach to complementing DPIAs with approved codes of conduct as a way to assess the impact of the processing operations performed by controllers or processors (Art. 40 and Art. 35.8), ADIAs should be complemented and further detailed in industry best practices, codes of conduct, codes of practice, standards and industry-led certification mechanisms." (p.19)

XIII. "The study of policy innovation starts from the proposition that there is no single universal best policy design, or best regulatory technology. Instead there are contextual criteria for success, which imply different regulatory designs for different problems, situations, societies, and institutional settings. We must test policy ideas, learn from empiricism, and adapt regulatory technology over time." Wiener 2004 (p.495)

XIV. "R]isk governance experts and scholars have developed new frameworks that continue to value scientific data but alongside other more qualitative measures of risk (…) In particular, these risk governance frameworks have three important features: (1) they focus on broadening participation in the risk governance process, including a range of key stakeholders; (2) they value qualitative data and policy analysis; and (3) they use deliberative, multi-stakeholder processes." Budish (2020), "AI & the European Commission's Risky Business", in https://medium.com/berkman-klein-center/ai-the-european-commissions-risky-business-a6b84f3acee0.

XV. This is similar to what was argued in the Oxford University's AI Governance Group's Response to the EC (2020) as an additional criteria for risk assessment: "Consider incorporating the scale of use (number of users, frequency of use) of a given AI application into the risk assessment procedure." (p.2)

XVI. For instance, models that recommend police coverage of neighborhoods based on past arrests will lead to an increase in arrests in those neighborhoods due to the increased coverage, data that will further influence the model, creating a feedback loop that could cause overpolicing unrelated to the actual rate of crime; models that deny loans to subjects that share a particular characteristic will never be able to 'learn' whether this assessment was accurate or should be adjusted; etc.

XVII. In the final Assessment List for Trustworthy AI (ALTAI), the HLEG advised that ALTAI is "best completed involving a multidisciplinary team of people. These could be from within and/or outside your organisation with specific competences or expertise on each of the 7 requirements and related questions." (p.4) The HLEG recommended identifying the following types of stakeholders to incorporate into the group that conducts the ALTAI, though gave no further detail about the types of responsibilities each of the roles should have within the ALTAI: AI designers and AI developers of the AI system, data scientists, procurement officers or specialists, front-end staff that will use or work with the AI system, legal/compliance officers, management.

XVIII. A prescriptive criteria may be useful as a presumption of potential high risk, but the actual high-risk determination should be made by organisations carrying out the risk assessment process. Along these lines, the Center for Information Policy Leadership's (CIPL) Response to the EC (2020) argued that "[p]roviding suggestive criteria, examples or presumptions of high risk would be of more practical use to those developing and using AI – including SMEs – than rigid ex-ante lists of high risk applications, as this is more suited to the highly contextual and evolving character of AI." (p.6)

XIX. A related argument stressing the importance of the procedural approach is the fact that a prescriptive approach seems to be ignoring long standing EU guidance on how to make proper risk evaluations. See Submission of the Centre for the Governance of AI, Future of Humanity Institute, University of Oxford's Response to the EC WHite Paper on AI (2020): *"[t]his [binary] approach does not seem to follow other existing risk assessment methodologies, which usually define risk in a given scenario as a function of the "combination of the probability of occurrence of a hazard generating harm in a given scenario and the severity of that harm,"* (p.3) citing to the EU General Risk Assessment Methodology.

XX. The need to incorporate benefits as mitigating factors into any risk analysis has been raised by many experts. See, e.g. Brookings Institute's Response to the EC White Paper (2020) on AI: "[R]egulation can […] raise barriers to the development and application of AI. This underscores the need for a balanced approach to AI regulation, one that takes into account the risks of AI and its benefits, a regulatory process informed by experts and science, that is sufficiently flexible to respond and learn from experiences with AI use-cases."

End notes

(p.5); Center for Information Policy Leadership's (CIPL) Response to the EC (2020): "High risks related to an AI system may be overridden by compelling benefits to individuals, organisations and society at large." (p.7). Thus organisations should be allowed to rebut presumptions of high risk AI by demonstrating countervailing benefits of the AI to individuals or society); United States (Proposed) Algorithmic Accountability Act (2019), authorizing the FTC to require companies to conduct ADIAs that would include assessment of all relevant benefits and costs; US OMB Memorandum to Federal Agencies of Regulation of AI (2020), which also emphasizes that executive agencies should factor in the benefits of AI in evaluating the risks entailed through use or application of the AI.

XXI. Canada's Directive on Automated Decision Making (2019), and its accompanying Algorithmic Impact Assessment tool that the relevant agencies can use to comply with the Directive, does something similar. The tool uses a scale of different "Impact Levels" (1 to 4 based on reversibility and duration of impact) to determine what statutory obligations apply to the agency or branch attempting to deploy or procure the AI (both in obtaining approval and after approval is obtained). The categories of potential obligations include the following: peer review obligations, public notice requirements, human in the loop requirements, explainability requirement for how decisions are made, testing requirements, monitoring requirements, training requirements, contingency planning requirements, and approval to operate requirements. Each different impact level is assigned (by the statute and the report that the tool generates) different types of obligations within each of those categories of requirements.

XXII. The importance of drafting the policy prototype, and in particular the playbook, in a way that is readily accessible, usable and actionable for technologists, opens exciting synergies and collaboration opportunities between experimental governance and the field of legal design. For context, legal design is an innovative approach that has gained traction over the last five years and is dedicated to rendering laws and regulations more understandable and easy to use for non-legal professionals. See, *e.g.* Minzoni 2020. We believe that legal design, as a discipline and set of methodologies, can help render the law – whether actual law or prototype law – more accessible and usable by non-lawyers, bridging language barriers between technologists and policy / lawmakers.

XXIII. This is very much in line with the increasing call to look beyond individual harms and pay attention to societal level harms in risk assessment. See, the EU High Level Experts Group on AI's Recommendations for Trustworthy AI (2020b) (establishing "societal and environmental harms" as one of the seven core requirements of AI, and incorporating that requirement into its final Assessment List for Trustworthy AI), German Data Ethics Commission Recommendations on AI (2019) (determination of the severity of harm must include potential harms to society and social cohesion, not just harms to individual rights), US OMB Memo to Federal Agencies of Regulation of AI (2020) *("Agencies should, when consistent with law, carefully consider the full societal costs, benefits, and distributional effects before considering regulations related to the development and deployment of AI applications."* (p.5)), IEEE 7010-2020 – Recommended Practices for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being (2020) (a comprehensive framework for AI risk analysis, expressly leaning more toward societal level impacts than individual impacts), Mantelero's (2018) "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment" (academic paper proposing a combination of human rights, ethics and social impact framework for AI risk analysis, attempting to incorporate societal risks in ways that more traditional AI impact assessments have not), Singapore and WEF Self-Assessment Guide for Organisations (2020) (wherein numerous of the assessment questions ask the organisation to consider potential impacts of the AI application on society collectively).

XXIV. As the National Institute of Standards and Technology (NIST) in the U.S. argued in its Plan for Federal Engagement in Developing Technical Standards and Related Tools (2019): *"While stakeholders in the development of this plan expressed broad agreement that societal and ethical considerations must factor into AI standards, it is not clear how that should be done and whether there is yet sufficient scientific and technical basis to develop those standards provisions."* (p.16)

XXV. In the final Assessment List for Trustworthy AI (ALTAI) (2020b), the HLEG suggested a number of questions to be asked, answered, and recorded before beginning the ALTAI. Some of those questions delve into the integration of DPIAs with AI Risk assessment processes: *"Have you put in place processes to assess in detail the need for a data protection impact assessment, including an assessment of the necessity and proportionality of the processing operations in relation to their purpose, with respect to the development,*

End notes

deployment and use phases of the AI system? Have you put in place measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system?" (pp. 5-6)

XXVI. A similar idea of putting these systems in practice in order to learn before acting is the one advanced by GoodAI (2020), which in their Response to European Commission proposes a risk monitoring process whereby firms deploying AI can voluntarily submit to collaborative and ongoing risk assessments at reference testing centers, giving the European Commission an opportunity to see risk assessment in practice and learn about the types of AI risk that it should be concerned with regulating.

Bibliography

Bason, Christian. *Design for policy*. Routledge, 2016.

Brest, Paul. "The power of theories of change." *Stanford Social Innovation Review* 8, no. 2 (2010): 47-51. http://sc4ccm.jsi.com/wp-content/uploads/2016/07/The-Power-Of-Theories-Of-Change.pdf

Brookings Institute. "Submission to the EC White Paper on Artificial Intelligence (AI) – The importance and opportunities of transatlantic cooperation on AI." (2020). https://www.brookings.edu/wp-content/uploads/2020/06/AI_White_Paper_Submission_Final.pdf

Brown, Tim, and Barry Katz. "Change by design." *Journal of product innovation management* 28, no. 3 (2011): 381-383.

Brown, Tim. "Design thinking." *Harvard business review* 86, no. 6 (2008): 84.

Budish, Ryan. "AI & the European Commission's Risky Business." (2020). https://medium.com/berkman-klein-center/ai-the-european-commissions-risky-business-a6b84f3acee0.

Calvo, Rafael A., Dorian Peters, and Stephen Cave. "Advancing impact assessment for intelligent systems." *Nature Machine Intelligence* 2, no. 2 (2020): 89-91.

Center for Democracy and Technology. "Response to the European Commission Consultation on the White Paper on Artificial Intelligence – a European Approach to Excellence and Trust Supporting Document." (2020). https://cdt.org/wp-content/uploads/2020/06/CDT-Supporting-Document-for-EU-Commission-Open-Consultation-on-the-AI-White-Paper-June-2020.pdf

Chair Legal and Regulatory Implications of Artificial Intelligence MIAI (Grenoble Alpes). "Consultation on the White Paper on Artificial Intelligence – A European Approach." (2020). https://ai-regulation.com/wp-content/uploads/2020/06/AI-Regulation-Submission-EU-AI-FINAL-Post%C3%A9.pdf

Centre for Information Policy Leadership (CIPL). "Response to the EU Commission White Paper On Artificial Intelligence – A European approach to excellence and trust." (2020). https://www.huntonprivacyblog.com/wp-content/uploads/sites/28/2020/06/cipl_response_to_eu_consultation_on_ai_white_paper__11_june_2020_.pdf

Considerati, ECP and Platform for the Information Society. "Artificial Intelligence Impact Assessment." (2018). https://www.considerati.com/static/default/files/documents/pdf/Artificial%20Intelligence%20Impact%20Assessment%20-%20English[2].pdf

Council of Europe. "Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems." (2020). https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154

Council of Europe. "Governing the Game Changer – Impacts of Artificial Intelligence Development on Human Rights, Democracy and the Rule of Law. – Conclusions from the Conference." (2019). https://www.coe.int/en/web/freedom-expression/aiconference2019

Data & Society. "Governing Artificial Intelligence Upholding Human Rights & Dignity". (2020). https://datasociety.net/library/governing-artificial-intelligence/

European Commission. "Proposal for a legal act of the European Parliament and the Council laying down requirements for Artificial Intelligence." (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=PI_COM:Ares(2020)3896535&from=NL

European Commission (EC). "White Paper on Artificial Intelligence – A European Approach to excellence and trust." (2020a).
https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Bibliography

European Commission (EC). High-Level Expert Group on Artificial Intelligence (AI HLEG). "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment." (2020b). https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

European Commission (EC). "Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC". (2020c). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en

European Commission (EC). "Ethics guidelines for trustworthy AI." High Level Expert Group on Artificial Intelligence. (2019a). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Commission (EC). "Communication: Building Trust in Human Centric Artificial Intelligence." (2019b). https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679." (2017). https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236

European Union (EU) Agency for Fundamental Rights. "Getting the Future Right – Artificial Intelligence and Fundamental Rights." (2020). https://eu2020-bmjv-european-way-on-ai.de/storage/documents/FRA-2020-AI-and-fundamental-rights.pdf

Facebook. "Facebook's Comments on European Commission White Paper on Artificial Intelligence – A European Approach." (2020). https://ai.facebook.com/blog/collaborating-on-the-future-of-ai-governance-in-the-eu-and-around-the-world/

Fuller, Lon L. "The morality of law." (1964).

German Data Ethics Commission. "Opinion of the Data Ethics Commission." (2019). https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf

GoodAI. "Response to the Consultation on the White Paper on Artificial Intelligence – A European Approach: Connecting Ecosystem of excellence and Ecosystem of trust: A concrete proposal (2020), paper available at the European Commission's White Paper on Artificial Intelligence consultation page, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12270-White-Paper-on-Artificial-Intelligence-a-European-Approach/public-consultation .

Government of Canada. "Algorithmic Impact Assessment." (2020). https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Government of Canada. "Directive on Automated Decision-Making." (2019). https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

IEEE. "IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being," in IEEE Std 7010-2020, vol., no., pp.1-96, 1 May 2020, doi: 10.1109/IEEESTD.2020.9084219. https://ieeexplore.ieee.org/document/9084219

IEEE. "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems." (2019). https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Information Commissioner's Office (ICO). "Big data, artificial intelligence, machine learning and data protection." (2017). https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Bibliography

Katyal, Sonia K. "Private accountability in the age of artificial intelligence." *UCLA L. Rev.* 66 (2019): 54. https://heinonline.org/HOL/LandingPage?handle=hein.journals/uclalr66&div=6&id=&page=

Kelsen, Hans. "The law as a specific social technique." *U. Chi. L. Rev.* 9 (1941): 75.

Kimbell, Lucy. "Introducing sprints to policymaking." *in* Researching design for policy. Findings from academic research in the UK Cabinet Office Policy Lab and beyond. (2015) (https://researchingdesignforpolicy.wordpress.com/2015/03/18/introducing-sprints-to-policymaking/

Kontschieder, Verena. "Prototyping in Policy: What For?!." (2018). https://conferences.law.stanford.edu/prototyping-for-policy/2018/10/22/prototyping-in-policy-what-for/

Leurs, Bas and Kelly Duggan. "Proof of concept, prototype, pilot, MVP – what's in a name? Four methods for testing and developing solutions." (2018). https://www.nesta.org.uk/blog/proof-of-concept-prototype-pilot-mvp-whats-in-a-name/

Mantelero, Alessandro. "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment." *Computer Law & Security Review* 34, no. 4 (2018): 754-772.

Minzoni, Marco. "Legal Design: How Design Can Make Law Easier to Understand." (2020). https://www.pixartprinting.co.uk/blog/legal-design/

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3, no. 2 (2016): 2053951716679679.

National Institute of Standards and Technology. US Department of Commerce. "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools Prepared in response to Executive Order 13859." (2019). https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf

OECD. "Recommendation of the Council on Artificial Intelligence." (2019) https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Office of Management and Budget (OMB). "Guidance for the Regulation of Artificial Intelligence Applications." (2020). https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

Personal Data Protection Commission Singapore (PDPC) and Infocomm Media Development Authority (IMDA). "Model Artificial Intelligence Governance Framework Second Edition." (2020). https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Position paper on behalf of Denmark, Belgium, the Czech Republic, Finland, France, Estonia, Ireland, Latvia, Luxembourg, the Netherlands, Poland, Portugal, Spain and Sweden on innovative and trustworthy AI. "Non-paper – Innovative and trustworthy AI: two sides of the same coin." (2020). https://www.permanentrepresentations.nl/documents/publications/2020/10/8/non-paper---innovative-and-trustworthy-ai

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." (2018). https://ainowinstitute.org/aiareport2018.pdf

Schiff, Daniel, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. "Principles to Practices for Responsible AI: Closing the Gap." *arXiv preprint arXiv:2006.04707* (2020). https://arxiv.org/abs/2006.04707

UNESCO. "Preliminary report on the first draft of the Recommendation on the Ethics of Artificial Intelligence." (2020). https://unesdoc.unesco.org/ark:/48223/pf0000374266

UNI Global Union. "Top 10 Principles for Ethical Artificial Intelligence." (2018). http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

Bibliography

University of Oxford. Future of Humanity Institute. "Consultation on the European Commission's White Paper on Artificial Intelligence: a European approach to excellence and trust." (2020). https://www.fhi.ox.ac.uk/wp-content/uploads/EU-White-Paper-Consultation-Submission-GovAI-Oxford.pdf
US Congress. "H.R.2231 – Algorithmic Accountability Act of 2019." (2019). https://www.congress.gov/bill/116th-congress/house-bill/2231

Villa Alvarez, Diana Pamela, Valentina Auricchio, and Marzia Mortati. "Design prototyping for policymaking." (2020).

Wiener, Jonathan B. "The regulation of technology, and the technology of regulation." *Technology in Society* 26, no. 2-3 (2004): 483-500.

World Economic Forum (WEF). "Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations." (2020).
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf

World Economic Forum (WEF). "Agile Governance Reimagining Policy-making in the Fourth Industrial Revolution." (2018). https://www.weforum.org/whitepapers/agile-governance-reimagining-policy-making-in-the-fourth-industrial-revolution

Annex
Automated Decision Impact Assessment (ADIA) prototype

ADIA prototype law

NOTE:
This document is to be used solely for the completion of Open Loop's Policy Prototyping Program: Automated Decision Impact Assessment. The sole purpose of this document is to elicit feedback on its content and format from the participating companies to the Policy Prototyping program. It is a fictional document deprived from any binding or legal normativity.

Recitals

Subject matter and scope

1. This Automated Decision Impact Assessment (ADIA) Prototype (hereinafter "ADIA", "Policy Prototype" or "Prototype") makes a distinction between automated decision-making systems (ADM) and processes that pose a low risk to the rights and freedoms of natural and legal persons and those that may pose a high risk, as defined herein.
2. This prototype shall not apply to the use of automated decision-making systems by natural persons in the course of a personal or household activity, with no connection to a professional or commercial activity. Where the automated decision-making system is provided by an actor to the natural person in the context of a professional or commercial activity, this Policy Prototype shall apply to them, as set forth herein.

Definitions

3. This Prototype refers to various actors in the field of automated decision-making systems, in particular developers, users, end-users, and subjects.
4. The developer is the natural or legal person who developed the automated decision-making system. This actor may only provide the learning algorithm, but it is more likely that they will be the person or organization responsible for selecting the (training) data and relevant learning algorithms and the subsequent creation and/or training of the model.
5. The user is the natural or legal person deploying an automated decision-making system to achieve a particular goal. Separately from the definition of what entails a high or low risk application, and generally speaking, this will be an organization, such as tax authorities detecting fraud, social media platforms providing automated personalized recommendations or banks assessing the creditworthiness of a client. The automated decision system deployed can be a stand-alone system, or an integral part of the delivery of a product or service.
6. The end-user is the natural or legal person who is intended to use the automated decision-making system, as opposed to actors involved in developing or determining its use. The end-user would be the actor informed by the decision of the automated decision-making system and/or doing the follow up of the automated decision. For example, a doctor getting advice on a treatment from an automated decision-making system, or a border control officer conducting a search of a person that was flagged by such a system. The end-user can be an employee of the user or independent of the user, using the automated decision-making system as a product or service of the user.

7. The subject is a natural or legal person that is directly or indirectly subjected to or impacted by an automated decision-making system.
8. While the actors each have a discrete role, in practice these roles might coincide. For example, with autonomous cars the passenger/occupant could be considered both the user, end-user and the subject. At the same time, the car manufacturer may also be considered the user.

Principles

9. For the purpose of this Policy Prototype a distinction is made between automated decision-making systems in general and those whose decisions likely pose a high risk to the rights and freedoms of natural and legal persons.
10. When developing and/or using an automated decision-making system the following principles should at a minimum be taken into account according to a risk-based approach: respect for fundamental rights and human agency, the need for human oversight, technical robustness, accuracy and safety, privacy and data protection, transparency and interpretability, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability. Where the use of automated decision-making systems may pose a high risk to the rights and freedoms of natural and legal persons, the user shall take specific technical and organizational measures to ensure the negative impact of these effects is minimized and the above requirements met.
11. Whether there is a high risk to the rights and freedoms of natural and legal persons must be judged on the context, nature, purpose, and scope of the application. There is a high risk when there is a significant chance that the automated decisions made by the automated decision-making system, or the subsequent actions taken by users, end-users, or subjects on the basis of that automated decision, result in negative effects with a significant adverse impact on the rights and freedoms of natural and legal persons.
12. Effects with a significant negative impact on the rights and freedoms of natural and legal persons may include loss of life or injury, financial or property damage, reputational damage and interference with fundamental rights such as the right to equality and non-discrimination, right to privacy, and the right to freedom of speech. In the context of automated decision-making, particular attention should be given to economic, psychological, and societal harms that may flow forth from automated decision-making. These include ***inter alia*** (legal) effects that lead to a loss of economic opportunity such as price discrimination, employment discrimination or unfair commercial practices; effects that lead to psychological harm such as self-censorship, loss of self-worth, and loss of personal autonomy; and collective harms such as a loss of liberty and economic or political instability. A taxonomy of harms can be found in the ADIA prototype guidance (playbook).
13. Automated decision-making systems should be used to augment human agency, increase human autonomy, and contribute to human well-being. Nonetheless, it is important to note that automated decision-making systems may limit human agency and may be used to influence, nudge, or manipulate end-users and subjects without their knowledge. Developers and users should take proper measures to avoid leveraging the persuasive capabilities of automated decision-making systems for unduly influencing or manipulating end-users and subjects.
14. Subjects have the right not to be subjected to automated decision-making without any meaningful human intervention when such a decision has a significant negative impact on their rights and freedoms.

15. The right to the protection against the consequences of automated decision-making is not an absolute right; such right must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality. For instance, an automated assessment of creditworthiness may be warranted when the subject wishes to enter into a contract and receive a credit.

Risk management and governance

16. The user shall assess whether the decisions made by automated decision-making systems may result in high risk to the rights and freedoms of natural and legal persons. If so, the user must conduct an impact assessment with regards to the automated decision-making system prior to deployment of the system. Particularly, this impact assessment will evaluate the risks that the decision-making poses to subjects, will consider the accuracy, quality and representativeness of the data used for the decision-making, the accuracy and quality of the (trained) models and the broader system of risk management for automated decision making. The impact assessment will propose measures to address the risks. After completing the impact assessment, the risk should be reduced to an acceptable level. An acceptable level of risk is defined as not having a significant adverse effect on the rights and freedoms of subjects. Adverse effects may be, but are not limited to, death, bodily harm, financial damage, reputational damage, discrimination, and stigmatisation. Where the user requires the assistance or input of the developer to assess and reduce the risks, it shall be the responsibility of the user to enlist the help of the developer.
17. Automated decision-making systems are systems that often change significantly throughout their life cycle. Therefore, the user must regularly update the impact assessment to ensure that the decisions that result from the system are still meeting the requirements set forth by this Prototype.
18. Where an impact assessment indicates that the decision-making process would (in the absence of safeguards, security measures, and mechanisms to mitigate the risk) have a high risk, and the user is of the opinion that the risk cannot be mitigated by reasonable means in terms of available technologies and cost of implementation, the supervisory authority should be consulted prior to the deployment of the automated decision-making system.
19. The use of automated-decision making on a large scale, affecting communities or society as a whole; use of automated decision-making leading to unfair bias and discrimination; use of automated decision-making limiting human agency; and use of automated decision-making in the context of surveillance may pose significant risks. For these high-risk areas, an impact assessment is in any case warranted. Further guidance is provided in the playbook.
20. To reduce the risk of automated decision-making, the user should have a robust system of risk management and governance. Given the potential impact of automated decision-making, the highest management should be involved in managing risk and ensuring a legitimate and ethical application of automated decision-making.
21. In order to demonstrate compliance with this Policy Prototype, the user should adopt internal policies and implement measures which meet the requirements set out in this Prototype. Risk management and governance should cover topics such as allocation of responsibility, policies and procedures, escalation protocols, data management, impact assessments, model management, subject rights, awareness raising, monitoring, oversight, and reporting. Furthermore, the developer and user are urged to set up mechanisms to facilitate and protect whistleblowers.

Chapter 1: subject matter and objectives

Article 1: subject matter and objectives

1.1 This Policy Prototype lays down rules to help protect the fundamental rights and freedoms of natural and legal persons that may be affected by automated decision-making.

1.2 This Policy Prototype lays down rules to help ensure a trustworthy application of automated decision-making.

1.3 This Policy Prototype aims to stimulate the development and use of automated decision-making for the well-being of society.

Article 2: material scope

2.1 This Policy Prototype shall apply to the development, production, distribution, and use of automated decision-making systems whose use may result in a high risk to rights and freedoms of natural or legal persons.

Chapter 2: Definitions

Article 3: definitions

a. 'Actor' means the developers, users, end-users, subjects, and any other party that contributes to the design, development, production, distribution, training, and/or deployment of automated decision-making systems and/or is affected by such a system or its decisions.

b. 'Algorithm' means a finite sequence of instructions or set of rules designed to complete a task or solve a problem.

c. 'Model' means the result of training an algorithm with training data. This model is a mathematical representation of the learned domain and is used to map inputs to outputs. The model is the primary component of an automated decision-making system.

d. 'Fully automated decision' means a decision made by an automated decision-making system which is acted upon without any meaningful human intervention.

e. 'Automated decision-making system' means a computational process derived from machine learning, statistics, artificial intelligence, or other data processing technique, that makes a decision or facilitates human decision-making.

f. 'Developer' means the natural or legal person responsible for the technical development of the automated-decision making system.

g. 'User' means the natural or legal person deploying an automated decision-making system to achieve a particular goal.

h. 'End-user' means the natural or legal person using the automated decision-making system for the purposes intended by the user.

i. 'Subject' means the natural or legal person subjected directly or indirectly to a decision of an automated decision-making system.

j. 'Automated decision impact assessment' (ADIA) means a systematic assessment of the impact of the envisaged automated decision-making system and its application.

Chapter 3: Risk management and governance

Article 4: Risk assessment

4.1 Prior to the deployment of an automated decision-making system, the user shall assess the risks of the envisaged automated decision-making system and its application on the rights and freedoms of natural and legal persons.

4.2 In those cases where the application of an automated decision-making system is likely to result in a high risk to rights and freedoms of natural or legal persons, the user shall carry out an automated decision impact assessment prior to the deployment.

4.3 An automated decision impact assessment referred to in paragraph 2 shall in any case be required in case of:

- potential unfair bias or discrimination towards subjects, including price discrimination, employment discrimination, or discriminatory differential access to services;
- potential loss of control or agency for the subject, including economic or psychological manipulation; or
- large scale application of automated decision-making, including profiling and systematic monitoring, that may affect communities or society as a whole;

4.4 An automated decision-making system impact assessment shall contain at least:

- a detailed description of the automated decision-making system, its design, its training, its data, and its purpose;
- an assessment of the quality, integrity and representativeness of the data used to train the underlying model;
- an assessment of the risks involved for natural and legal persons, with a specific focus on subjects and for end-users; and,
- the measures envisaged to address the risks including safeguards, security measures and mechanisms protecting the rights and freedoms of end-users and subjects and to demonstrate compliance with this Policy Prototype, taking into account the rights and legitimate interests of those concerned.

4.5 In those cases where the automated decision impact assessment indicates that the application may result in a high risk to the natural rights and freedoms of natural and legal persons and these risks can or will not be mitigated, the user shall prior to the deployment consult with the supervisory authority.

Article 5 Governance

5.1 Developers and users shall have adequate and effective internal governance structures and measures in place to ensure robust oversight on their respective role or roles in the design, development, deployment, and training of automated decision-making systems.

5.2 The highest management within the relevant organization shall on an on-going basis be involved in, and responsible for, explicating the ethical values that guide the process of design, development, deployment, and training of automated decision-making systems.

5.3 Developers and users shall have a sound system of risk management and internal controls in place, specifically aimed at identifying, assessing, documenting, and addressing the risks involved in the design, development, deployment, and training of automated decision-making systems. Such measures include establishing adequate monitoring and reporting schemes.

5.4 The developer and users are able to demonstrate that the measures taken are adequate to mitigate the risk posed by the automated decision-making system used.

ADIA Prototype Guidance / Playbook

(the playbook would complement forthcoming legislation and could be the basis of soft law instruments: codes of conduct, codes of practice, standards, certifications, industry guidelines, etc)

In this section we set out ways to comply with the proposed Policy Prototype. By implementing the elements from the playbook, an organization is in a good position to comply with the prototype law.

Risk assessment

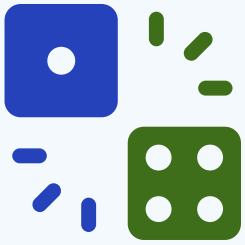
The prototype law requires that an automated decision-making impact assessment be completed for automated decision-making systems that may pose a high risk to natural and legal person's rights and freedoms.

Quantifying risk

The impact of the ADM system on people or society will differ per application. To determine whether there is a high or low risk, consider this (industry accepted) formula:

Risk = probability that a disruptive event occurs x severity (the negative impact) of that event.

When developing or deploying an ADM system, the developer and user need to assess, based on their concrete application, the probability that a disruptive event will lead to a negative impact, as well as the severity of the impact.



1) Probability
Different factors may play a role in the probability of a disruptive event occurring, such as the type of application, the scope of the application, the number of subjects involved, the complexity of the model, the novelty of the domain where the application is being deployed, the investment made in issue spotting, and the robustness and reliability work done in the building and testing phase. These factors need to be weighed based on the circumstances of the case.

2) Impact
The impact of a disruptive event on natural and legal persons coming from an ADM needs to be determined on a case-to-case basis, taking into account the scope, nature and context of the application. Negative impacts commonly associated with ADM have been described in literature, see: “Taxonomy of potential harms,” “Table 2: Example list of harms that may have significant effect on natural and legal persons.”

Quantifying risk in practice

In order to quantify risk, developers and/or users should implement a risk assessment process. Such a process could look something like this:

Step 1: Describe the proposed ADM system

Describe the ADM system and its goal. First, determine the context in which the ADM system will be used. The goal of the system and the context in which it operates determines to a large extent the potential risk of the system. For example, ADM systems used in a medical context to determine the type and amount of medicine to administer will pose a greater risk to people than a recommendation engine for movies on a streaming platform.

In describing the ADM system, explain how the system works, what stakeholders are involved and how the stakeholders interact (socio-technical context).

Step 2: Assess how ADM changes the existing situation

The second step is to determine for a given context whether ADM introduces new risks or benefits, or changes the existing level of risk. To this end, the developer or user needs to determine how automating an existing human decision-making process, or introducing a new automated decision-making process, introduces new risks, or changes the existing level of risk for all stakeholders.

Step 3: Analyse the root cause of the change

Find the root cause for this change (for example, by introducing ADM, it became less clear why a decision was made. This opacity can be attributed to the complexity of the model). Use the list of potential harms (Table 2) to find relevant root causes that may lead to risks to the rights and freedoms of (in particular) subjects.

Step 4: Determine impact on stakeholders and associated values

Determine how the identified changes affect stakeholders (both positive and negative) and associate these changes with fundamental human rights (e.g. a hypothetical ADM process increases the risk of discrimination, thus undermining fairness and equality). Use recitals 11 and 12 as well as the taxonomy of harms from table 2 to determine potential negative impact of the proposed ADM system at the individual and societal level. Relate these to values as described in table 1.

Step 5: Determine value tensions

Discover any tensions between the values and their relation to stakeholders (e.g. "our algorithm nudges people into doing microtransactions on our platform improving our bottomline, but this may affect their autonomy and material well-being"; (another ***e.g.***) "proactively removing hate speech with AI reduces harm overall and is beneficial, but we can't have perfect AI accuracy in this task, and as a result some valid speech may mistakenly be taken down – this may affect some groups of people more than others"). Use the overview of values from table 1 and the taxonomy of harms (table 2) for this evaluation.

Step 6: Determine probability of negative impact occurring

Determine the likelihood of the negative impact manifesting itself. Devote specific attention to the decision-making model.

Step 7: Identify possible changes and mitigating measures

Identify possible changes to the design, alternatives, and/or risk mitigating measures that will reduce the negative effects for stakeholders. See "Table 3: Example List of Technical and Organizational Measures as Mitigating Measures" for help with this step.

Step 8: Assess consequences of changes and mitigating measures

Determine how these changes will affect the other stakeholders (both positive and negative).

Step 9: Decide which changes and mitigating measures to implement

Decide which changes to make based on established AI principles (your own and/or others you refer to), taking into account laws and norms.

Step 10: Implement and document

Implement and document the changes.

A developer and/or user could make a distinction between a *quick scan risk assessment* or ***pre-screening*** to determine whether a full scale automated decision-making impact assessment is necessary (stopping at step 6), or the user can follow and document all 10 steps and do a *full scale automated decision impact assessment*.

Overview of values relevant to AI

As the debate on the influence of AI/ML on our society is still in its infancy and concrete harms are very context dependent, the discussion generally takes place on the level of values. To discuss harms, connections to underlying values are often made. For instance, to discuss the harm of algorithmic manipulation, we refer to the shared value of personal autonomy.

Table 1: Overview of values relevant to AI

Value	Description	Possible relevance in the context of AI/ML and ADM
Privacy & Data Protection	Right to protection of a personal sphere and protection of personal data.	AI/ML systems are able to gather, process and infer data with an unprecedented scale and speed.
Personal autonomy	Ability of a person to decide what is good and bad for them. Ability to think and act without reliance on others and without their control or influence.	AI/ML systems enable (un)conscious delegation of personal autonomy. AI/ML systems enable advanced and personalized persuasion.
Human dignity	The notion that each person has an intrinsic worth that should be respected by other actors.	AI/ML systems enable datafication of persons. People may believe they are engaging with another person, rather than an automated system.
Liberty/Freedom	Ability to think and act without interference from others (negative liberty). Ability for self-realisation of the individual (positive liberty).	AI/ML systems can enable surveillance on an unprecedented scale in both the public and private sector.
Fairness	Equal and just distribution of benefits and costs (substantive fairness), acting fairly in decision-making (procedural fairness).	Potential for AI/ML systems' fairness to be impacted by replication of unfairness in society. Potential for AI/ML systems to be procedurally unfair if they are inaccessible and/or cannot be contested.
Responsibility	Moral obligation to act in a particular situation. Duty of care.	Responsibility may be 'outsourced' to AI/ML applications.
Accountability	Obligation to account for your activities, take responsibility for these actions and disclose them in a transparent manner.	AI/ML may make it harder to hold actors accountable for actions.
Democracy	Systems of governance based on self-rule by the people through chosen representatives, respect for the rule of law and human rights.	AI/ML systems may disrupt democratic processes by changing the flow of information and interactions of people.
Rule of Law	Governance by law, non-arbitrary exercise of (government) rule and power.	Potential for civil rights laws to be implicated by AI/ML systems.
Material well-being	Ability to derive well-being from material assets.	AI/ML systems may have micro-economic effects through *e.g.* personalized pricing. AI/ML systems may have macro-economic effects through automation and shifting of economic power.
Transparency	Condition that enables openness, honesty, visibility and accountability.	AI/ML systems may be opaque, undermining transparency and interpretability of decisions.

Taxonomy of potential harms

To make the abstract definition of a 'harmful or consequential decision' more tangible, we need to first describe and classify a number of harms that are of particular relevance in the context of AI/ML and ADM.

We can make a distinction between the individual dimension of a value and the collective dimension of a value. Conversely, harms can also have their effect on the individual and the collective level. While there is a strong focus on the individual dimension of values and harms in the literature, collective values and interests are also taken into account.[35]

Please note that the impact of an automated decision-making system is very much dependent on the context in which it is used. For each context, those who deploy automated decision-making systems should assess which individual and collective harms are relevant to consider.

See on the following pages an example list of harms that may have a significant effect on natural and legal persons.

35. Future of Privacy Forum, 2017. Unfairness by algorithm: Distilling the harms of automated decision-making

Table 2: Example list of harms that may have significant effect on natural and legal persons

Individual dimension of harm				
Economic harms				
Root cause	**Effect**	**Potential harm**	**Values at stake**	**Examples**
Prices can be tailored to groups and individuals based on their data.	Differential pricing	Price discrimination	Fairness, material well-being	Subjects get a higher price for a product because frequenting a product page and liking posts about the product signals strong interest.
Based on data of the target group and other insights, the target group can be influenced without their knowledge or consent.	Nudging	(Economic) manipulation of subjects	Fairness, personal autonomy, material well-being	Micro-targeting is used to show users more ads and status updates on a given topic to pivot them towards displaying the desired behaviour (e.g. healthier lifestyle).
AI empowers those employing it versus those who are subjected to it.	High power differential and information asymmetries	Unfair commercial practices (e.g. manipulation, hidden or false advertisement)	Fairness, autonomy, material well-being, human dignity	An employer evaluates employees based on monitoring unbeknownst to them.
The automatic decision-making process is intentionally designed to be biased towards certain groups and individuals.	Discrimination (intentional)	Loss of economic opportunity (e.g. employment discrimination), narrowing of choice	Fairness, material well-being, human dignity	Hiring algorithm is intentionally trained to favour young white males.
Data may be unknowingly discriminatory leading to groups and individuals being treated differently.	Discrimination (unintentional)	Loss of economic opportunity (e.g. employment discrimination), narrowing of choice	Fairness, material well-being, human dignity	Hiring algorithm unintentionally and disproportionately benefits young white males, based on training data that reflects societal/historical hiring patterns.
Data is used to profile subjects and place them into different target groups.	Differential access to goods and services	Narrowing of choice	Fairness, well-being	Credit scoring system used to differentiate between valuable and non-valuable customers.

Individual dimension of harms				
Psychological harm				
Root cause	**Effect**	**Potential harm**	**Values at stake**	**Examples**
Increased surveillance in the personal life.	Chilling effect	Self-censorship	Autonomy, human dignity, privacy	Not sharing information out of fear for future (currently unknown) consequences.
Opaque automated decisions.	Opacity, loss of understanding of decision	Loss of self-worth and self-efficacy	Human dignity, personal autonomy	Person is denied a job 'because the computer said so'.
Profiling is based on inaccurate data or models.	Unfair profiling	Stigmatization and reputational damage, loss of self-efficacy	Human dignity, personal autonomy, fairness	Innocent person is flagged as a terrorist based on parameters such as religion (false positives)
Decisions are delegated to ADM systems.	Humans no longer make the final decision	Loss of control	Personal autonomy, human dignity	Doctor is told by an ADM which actions to perform on a patient, undermining professional judgement of the doctor.
Hyper effective personalized persuasion.	Loss of control and dependency	Loss of self-worth and self-efficacy, dependency / addiction	Personal autonomy, human dignity,	Quantified self apps acting as personal coaches persuade people to make lifestyle choices (when to sleep, what to eat, when to put their phone away).
Decisions are made using inadequate data.	Incorrect decisions	Unfair decisions	Fairness, human dignity	Person is denied a loan based on incomplete or incorrect data regarding their financial situation.
The automatic decision-making process is biased towards certain groups and individuals.	Discrimination (intentional and unintentional)	Unfair decisions, discrimination and stigmatization	Fairness, human dignity, material well-being	A hiring algorithm for C-level functions intentionally/ unintentionally and disproportionately benefits white men based on societal/historical patterns, while exacerbating women's current underrepresentation in these functions.
Presenting the current situation as what it should be.	Constraint conception of future	Loss of creativity and reflexivity	Personal autonomy, human dignity	Because the model is trained on historic data it may perpetuate and/or strengthen the status quo and not think outside of the box.

Collective harms				
Root cause	**Effect**	**Potential harm**	**Values at stake**	**Examples**
AI increases the ability to surveil the public sphere.	Increased surveillance	Loss of liberty and autonomy, invasion of privacy, chilling effects	Liberty, personal autonomy, privacy	Use of facial recognition and emotion detection in closed-circuit television (CCTV).
Machine learning models are black boxes because the model is too complex to understand.	Opacity, loss of understanding of decision	Lack of understanding of a judgment, inability to verify correctness / accuracy, dehumanisation, bias and discrimination	Procedural fairness, autonomy, transparency	Organisations are unable to explain the rationale of risk classification of credit applications.
Based on data of the target group and other insights the target group can be influenced without their knowledge or consent.	Nudging	Manipulation of subjects	Fairness, liberty, autonomy, dignity	Election manipulation through micro-targeting of voters, targeting them with messages they are susceptible to.
Data may be used to profile subjects and place them into target groups.	Individual tailored approach	Loss of collectivity (e.g. in insurance)	Collectivity, fairness (equality)	Personalized insurance based on driving behaviour.
Prices are tailored to groups and individuals based on their data.	Differential pricing	Price discrimination	Fairness, material well-being	Vulnerable groups are targeted with higher interest rates for loans.
AI profiles perpetuate and/or strengthen existing societal biases.	Stigmatisation and stereotyping	Polarisation and division	Fairness (equality), Rule of law, human dignity	Sentencing algorithm incorrectly judging people of color to have a higher change of recidivism than white people.
AI is used to filter/moderate access to content.	Filter bubbles	Polarisation and division, loss of freedom of expression	Rule of law, democracy, material well-being	Hyper-personalized content on social media platforms strengthen existing opinions.
AI is used to doctor images, audio and video (deepfakes).	Manipulation, disinformation / fake news	Polarisation and division, political instability	Democracy, material well-being	Use of deepfakes to make it look like a person has said or done something they haven't.
AI empowers those employing it versus those who are subjected to it.	High power differential and information asymmetries	Inequality, stratification of society, political instability	Rule of law, equality	Employers have access to AI and ADM to manage and control the workforce, whereas the employees don't have access to the same systems.
The automatic decision-making process is biased towards certain groups and individuals.	Discrimination (intentional)	Discrimination, (political) instability	Rule of law, equality	An algorithm is used to assess the ethnicity of a person based on their name and intentionally exclude them from a service.
Data may be unknowingly biased, leading to groups and individuals being treated differently.	Discrimination (unintentional)	Discrimination	Rule of law, equality, material well-being	COMPAS algorithm incorrectly judged people of color to have a higher chance of recidivism than white people.
Data may be used to profile subjects and place them into target groups.	Differential access to goods and services	Discrimination, stratification of society	Fairness (equality), material well-being	An algorithm may exclude 'high risk' categories of people (e.g. those with debts) from services.

Mitigating measures

Where an automated decision-making system makes decisions that are likely to result in a high risk to rights and freedoms of natural or legal persons, the user should take the necessary technical and organizational measures to ensure that the automated decision-making system is developed and used in a lawful, ethical and robust manner.

The technical and organizational measures as described above shall in particular be aimed at ensuring that:

a. the automated decision-making system and the automated decisions are transparent and interpretable;
b. the automated decision-making system is technically robust, accurate, reliable and otherwise safe;
c. the automated decision-making system functions without unfair or unevenly distributed bias, including where the content of the training data reflects the diversity of natural or legal persons the decision-making is centered around;
d. the automated decision-making system is designed and functions in accordance with applicable data protection principles, including but not limited to purpose limitation, data minimization, limited storage periods, data quality, data protection by design and by default and data security;
e. the automated decision-making system is not used to unduly influence or manipulate the end-user or subject;
f. all actors have appropriate procedures in place ensuring accountability in accordance with this Regulation.

There are a number of different interventions that a user should take in order to mitigate the risks identified. These are set out in Table 3 and further described below.

Table 3: Example List of Technical and Organizational Measures as Mitigating Measures

	Typical activities	Activities that may contribute to trustworthy AI
Problem definition	• Specify Intended use case • Determine model specification	• Do AI risk assessment
Data selection / collection / preparation	• Select data • Collect data • Data preparation (cleaning, transformation, reduction, integration) • Feature engineering • Splitting data (training, validation, holdout)	• Determine whether data is representative for the problem / domain • Screen data for bias • Review feature engineering for risk • Ensure data protection
Model data	• Determine model evaluation criteria • Training candidate models • Model tuning • Model validation • Model selection and testing • Document modelling process	• Assess model evaluation criteria from the perspective of values (e.g. precision and recall trade-offs) • Determine presence of bias • Assess choices in model tuning and selection from the perspective of the subject • Ensure proper validation and testing
Interpret model outcomes	• Interpret model outcomes	• Ensure global or local interpretability of model. • Detect unwanted or unfair outcomes based on individual decisions • Perform fairness testing • Do an external assessment of outcomes (e.g. 3rd party audit)
Model deployment	• Document model training and testing process • Communicate model operation • Acceptance testing	• Revisit AI risk assessment, evaluate socio-technical interaction • Train end-users in the interaction with the model • Disclose use of ADM to end-users and subjects in particular • Provide subject rights and redress mechanisms
Monitoring & Enforcement	• Monitor performance • Implement feedback loop	• Monitor performance over time for degradation and bias • Monitor exception handling • Provide mechanisms for corrigibility and interruptibility • Periodically update AI risks assessment

Problem definition

The first step to reducing the potential risk associated with ADM is to assess the potential risk of the application in the conceptual phase, preferably using the methodology described above.

Data selection, collection, preparation

When selecting, collection and preparing data, users should assess the quality of the data. Users should in particular asses:

- The accuracy of the dataset, in terms of how well the values in the dataset match the true characteristics of the entities described by the dataset. In other words: how closely does the data represent reality?
- The accuracy of the dataset, in terms of trustworthiness of the data. In other words, have the data been gathered from a reliable source and can we trust that the values are accurate?
- The completeness of the dataset, both in terms of attributes and items.
- How recently the dataset was compiled or updated.
- The relevance of the dataset and the context for data collection, as it may affect the interpretation of and reliance on the data for the intended purpose.
- The integrity of the dataset that has been joined from multiple datasets, which refers to how well extraction and transformation have been performed.
- The usability of the dataset, including how well the dataset is structured in a machine-understandable form.
- Human interventions, e.g. if any human has filtered, applied labels, or edited the data.

Furthermore, users should be able to account for the data used in the AI/ML process (both the training data and subsequent input data). The user must be able to attest where the data came from, how it was used, how it wastransformed, etc. Therefore, users should properly log different data sources and in any case describe the steps in preparing the data.

In collecting and using data, personal data protection rules should be taken into account. Where possible, synthetic data should be used for training and testing purposes.

Model data

A very important element of any evaluation of an ADM system is the accuracy of the predictions. In determining accuracy for ADM systems that make decisions that impact people, classification accuracy (i.e. the total number of correct predictions divided by the total number of predictions) should not be considered an appropriate metric. A proper metric should for instance also take into account the effect of false positives and false negatives and account for choices made in the trade-off between recall (no false negatives) and precision (no false positives). Users should assess the risks for subjects associated with being classified as a false positive or a false negative and factor this into the recall/precision trade-off for the model.

Example recall-precision trade-off
In breast cancer screening, you want to ensure that you do not miss any possible tumors, therefore the recall should be set very high, even if this means more false positives (i.e. healthy women getting the wrong diagnosis). In fraud detection the situation may be different. If the recall is very high, a lot of people will be flagged as potential fraudsters, with negative consequences for them. In this situation it might be preferable to increase the precision. While you might miss out on some fraudsters, this might be an acceptable trade off if that means avoiding trouble for a lot of innocent people.

A specific part of accuracy often mentioned in the context of AI/ML is bias. Bias in decision-making could result, amongst other reasons, from using training data which contains (unknown) bias. This can for instance be attributed to e.g. sample selection errors, but also the fact that the training data accurately reflects a real world discriminatory situation. In order to avoid bias, there are ongoing conversations about removing sensitive attributes that are related to bias (e.g. gender or ethnicity). These methods are called blindness methods, however, when there are latent variables related to these sensitive attributes, bias may still occur. Therefore, users should consider using bias-aware approaches whereby the bias is captured in the model, but can be subsequently corrected after detection. Finally, users should follow the state of the art in terms of training, testing and validating models.

Interpret model outcomes

In the process of model selection, testing, and validation, understanding of the outcomes is important, if only to assess whether the model is performing adequately. But also from the perspective of the end-user, subject, and supervisory authority, explainability of model outcomes is important. Furthermore, the model should be tested for fairness; for instance, through counterfactual fairness testing. Counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. When the results are different based on for instance ethnicity, the model is almost certainly unfairly biased.

Model deployment

When finally deploying the model the user should train its end-users in the proper operation of the ADM system, in particular teaching them the scope, limitations, strengths, and weaknesses of the system. Furthermore, the user should observe how end-users and subjects interact with the system and determine if this is within the bounds of the original risks assessment. Therefore, it is also worthwhile to revisit the initial risk assessment to determine if it is still representative of the final outcome in this phase.

Monitoring and enforcement

Once the model is in operation, it is important how well it behaves over time. The outside world may change, making predictions based on the old reality less accurate over time. Therefore, the user should implement mechanisms to deal with model degradation. Also important is responding to subject feedback and complaints, as they may be indicative of issues with the underlying decision-making model.