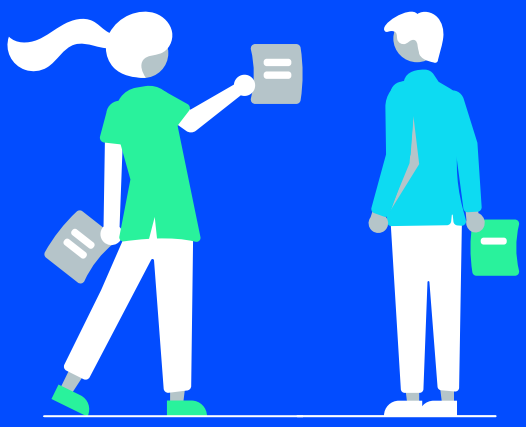
FROM POLICY TO PRACTICE

PROTOTYPING THE EU AI ACT'S TRANSPARENCY REQUIREMENTS



Knowledge Centre Data & Society

Artificial Intelligence Flanders

Citation: T. Gils, F. Heymans and W. Ooms (Knowledge Centre Data & Society), "From Policy To Practice: Prototyping The EU AI Act's Transparency Requirements", January 2024

Contact: thomas.gils@kuleuven.be or frederic.heymans@vub.be

www.data-en-maatschappij.ai

1. ABOUT THE KNOWLEDGE CENTRE DATA & SOCIETY

The Knowledge Centre Data & Society (KCDS) is the central hub in Flanders for the legal, social and ethical aspects of data-driven and AI applications.

The Knowledge Centre brings together knowledge and experience on this topic tailored to industry, policy, civil society and the general public. Specifically, our objectives include:

- **Disseminating information and knowledge** on the ethical, legal and social aspects of data-driven applications and AI. All publications are made publicly available and aim to create a positive and proactive effect between these innovations and our society.
- **Promoting structural initiatives** that strengthen vision development and valorise the social and economic opportunities of data-driven applications and AI among governments, industry and other social actors.
- **Stimulating public awareness** and debate on the benefits & drawbacks and the social, ethical and legal aspects of data-driven applications and AI, in all layers of society.
- **Building and supporting a network and learning environment** for stakeholders and strengthening collaboration between different policy levels and actors.
- **Contributing to the development of legal frameworks and guidelines** on the use and framing of AI and data-driven applications for policy makers, businesses, organisations and employees. Our policy prototyping project is one of the activities that we develop in order to achieve this objective.

Please visit our website for more information about the KCDS, our objectives and our offering.

TABLE OF CONTENTS

1. About the Knowledge Centre Data & Society 3

2. Executive summary 5

3. Introduction 7
3.1. Introduction to Policy Prototyping 7
3.2. Policy Prototyping at the KCDS 8

4. Policy Prototyping: methodology and process 9
4.1. Preparatory phase 10
4.2. Call for participants 10
4.3. Phase I 11
4.4. Phase II 13
4.5. Phase III 13
4.6. Phase IV 14

5. The AI Act - Text of Art. 13 & 52 AI Act 15
5.1. Article 13 AI Act (European Commission and Council) 15
5.2. Article 52 AI Act (European Commission and Council) 16

6. Results 17
6.1. Introduction 17
6.2. Prototype Instructions for Use 17
6.3. Prototype Disclaimers and decision-making processes 26

7. Feedback on AI Act 32
7.1. Feedback on Art. 13 AI Act 32
7.2. Feedback on Art. 52 AI Act 37
7.3. General feedback 40

8. Conclusion 41

9. Acknowledgements / Participants 42

2. EXECUTIVE SUMMARY

The European Union's (EU) AI Act emphasizes the importance of transparency regarding AI systems and imposes related requirements in order to foster trust and accountability. More specifically, the AI act (European Commission and Council-text) contains two sets of transparency requirements. The first set targets high-risk AI systems and requires to draft Instructions For Use (IFUs) for such systems (art. 13 AI Act). The second set targets "certain AI systems", including interactive AI systems and AI-generated/deep fake content, and requires them to disclose the artificial nature of the interaction or content (art. 52 AI Act).

The purpose of this policy prototyping project was to test these requirements by performing a mock compliance exercise and gather stakeholder feedback on the requirements. This report presents the results of this project and provides insights to policymakers and professionals. It illustrates how IFUs and disclaimers may look in practice and identifies lessons learned and best practices that should be taken into account when drafting these documents. In addition, the report provides legal feedback for policymakers regarding articles 13 and 52 AI Act. The prototype compliance documents are included in a separate annex to this report.

The report starts with an introduction to policy prototyping (part 3) and outlines the course and different phases of this project (part 4). Then, it discusses the prototype IFUs and disclaimers that were developed and the respective stakeholder feedback (part 6). The final section contains detailed legal feedback on articles 13 and 52 AI Act (part 7).

KEY FINDINGS RELATED TO THE PROTOTYPE IFUs AND DISCLAIMERS

Below, we highlight some of the findings in relation to the prototype IFUs and disclaimers, based on participant feedback.

Findings on the Instructions For Use

- When drafting IFUs, the primary target audience (i.e. the specific professional users) should be kept in mind at all times.
- The documents should have a logical structure, use simple, concrete and clear language tailored to the target audience and, use and adapt the design to enhance understanding.
- The documents should include extensive and explicit information regarding accuracy, performance and other relevant metrics.
- Information regarding input, training, validation and/or test data should be sufficiently detailed.
- The documents should include a minimum of installation and usage instructions.

Findings on the Disclaimers

- Disclaimers should establish a desired level of transparency, taking into account the potential target audience and accessibility considerations. Participants positively appreciated receiving more than the strictly legally required information.

- The disclaimer and related documentation should be accessible to different users, taking into account users with disabilities.
- Employing a layered approach in the structure of the disclaimer and related documentation can optimize user engagement and help avoid information overload.
- Disclaimers could include a provision for complaints and feedback to enhance transparency and user trust.

KEY FINDINGS RELATED TO THE TRANSPARENCY REQUIREMENTS OF THE AI ACT

Participants underline that compliance with both sets of transparency requirements (art. 13 and 52 AI Act) will require effort, proactive thinking and multidisciplinary experience and knowledge. This latter factor will often make it less straightforward for smaller providers of the respective AI systems to deal with the requirements due to lack of available financial and human resources. It is therefore unsurprising that respondents ask for sufficiently concrete guidance, templates and examples to facilitate the implementation of the transparency requirements and level the playing field.

A large majority of participants believe that both sets of requirements are desirable and support their inclusion in the regulatory framework. Transparency is seen as an essential tool to create trust. At the same time, participants provide many suggestions for amending both articles. A remarkable observation in that regard, is that participants wish to see additional requirements regarding the disclosure of technical details in both IFUs and disclaimers. Other respondents warn against requiring too much additional information as this may contribute to information overload and make AI systems more vulnerable. Finally, it is also argued that policymakers should keep legislation and policies up to date with technological changes in order to maintain the desirability of this type of requirements.

Finally, stakeholder feedback indicates that both provisions contain many unclear concepts that require clarification. Especially, the use of phrasing such as 'sufficiently transparent', 'reasonably foreseeable misuse' or 'appreciably resembles' confuses participants. It also follows from the feedback that a technical AI background appears to be an advantage in understanding art. 13 AI Act, while it is less so with regard to art. 52 AI Act.

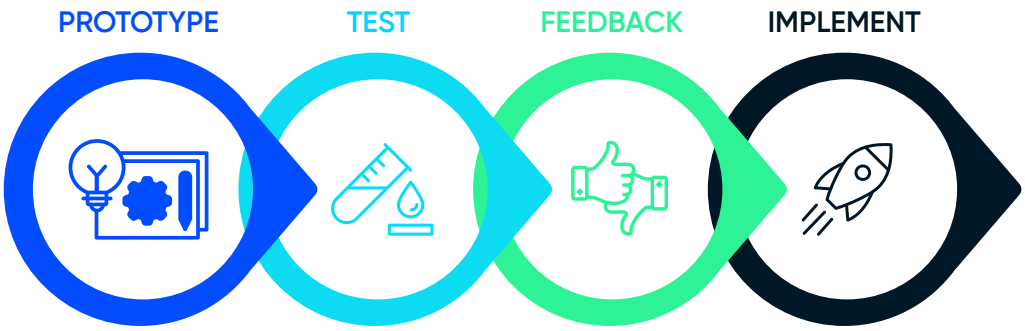
KEY FINDINGS RELATED TO POLICY PROTOTYPING

Throughout this project, the concept of policy prototyping has garnered positive feedback. Many participants recognized the significant value that policy prototyping may bring, thereby emphasizing the usefulness of exploring the application of regulatory requirements and provisions on an explicit fact-based use case. Participants agreed that this method can add much value to the policy implementation process. The positive reception underscores a broader consensus emphasizing the importance of actively involving a diverse array of stakeholders in the policymaking process. By gathering comprehensive insights from these different perspectives, policymakers can ensure a solid foundation for the policy implementation process

3. INTRODUCTION

3.1. Introduction to Policy Prototyping

Policy prototyping refers to a novel way of policymaking, comparable to product or beta testing. It can be understood as a form of user-centred policy design or applying the design thinking methodology to the legislative or policymaking process. Policy prototyping should enable policymakers to map the effects, strengths and limitations of a proposed policy and lead to more effective and evidence-based policymaking while avoiding the societal costs of 'bad policy'. Typically, a policy prototyping project consists of multiple phases:



- **Prototype:** prototyping implies the creation of basic models or designs for a machine or other product to test an idea or a concept in practice. In this context, prototyping entails drafting a new policy or law. Such prototypes can be elaborate or minimal, allowing to test specific features and find out 'what works' through several iterations.
- **Test:** a group of stakeholders performs a mock compliance exercise and implements the envisaged legal requirements.
- **Feedback:** participants provide feedback in relation to the mock implementation of the policy prototype.
- **Implement:** this feedback is used to evaluate if the law is effective and 'fit for purpose' and to complete and/or amend it accordingly, issue additional guidance, ...

In summary, policymakers and stakeholders can create tangible and practical prototypes of proposed policies and related compliance documents using this approach. These prototypes allow them to test and refine the policy measures before committing to a full-scale implementation.

Policy prototyping can help identifying potential gaps, challenges, or unintended consequences in an early stage of the policymaking process. It enables policymakers to make necessary adjustments and improvements to the policy, and stakeholders to prepare for future policy. In essence, policy prototyping may bridge the divide between policy design and actual implementation, enhancing the effectiveness, feasibility and acceptance of policies while minimising the risk of unanticipated policy mistakes or failures.

At the same time, policy prototyping projects should also consider some possible concerns for which they should ensure transparency or accountability. More specifically, the group of participants involved in a project ideally reflects the diverse group of stakeholders affected by the envisaged policy, while public transparency regarding the participants also needs to be ensured. Additionally, policy prototyping projects will generally be conducted with small testing groups. This may lead to casuistic results, reducing their representativity and scalability, as the results may not be applicable on the large scale on which regulation usually applies.

In part 4, we will explain in more detail how we applied this approach (including the concerns) in the policy prototyping project which is the subject of this report.

If you wish to have more in-depth information on policy prototyping and other initiatives in this area, we refer to the blog "Design thinking in the legislative process: the key to useable legislation?" written by KCDS-researchers in 2021 [1].

3.2. Policy Prototyping at the KCDS

In 2022, the KCDS conducted a first policy prototyping experiment. This experiment focused on the **scope of application of the EU AI Act** and the lists of prohibited and high-risk AI systems, as included in the European Commission proposal [2]. The main conclusions and results can be found in the following online publication: Policy Prototyping: an assessment of Articles 5 & 6 of the EU AI act [3].

After evaluating our first experience with policy prototyping, we decided to organise a second initiative, following a modified approach with increased stakeholder involvement and interaction. Our current project focused on the **EU AI Act's transparency requirements**. More precisely, it concerned the transparency requirements for high-risk AI systems (art. 13 AI Act) and the transparency requirements for "certain AI systems", including interactive AI systems, emotion recognition and biometric categorization and AI-generated/deep fake content. (art. 52 AI Act)

In the course of this project, we pursued four objectives:

1. **Examine** the envisaged transparency requirements in detail
2. Create **operational guidance** that includes prototype instructions of use for high risk AI systems (under art. 13 AI Act) and prototype disclaimers (under art. 52 AI Act)
3. Gather **feedback** on these transparency requirements and their applicability, feasibility, understandability
4. Provide our **findings and lessons learned** to policymakers and other stakeholders.

[1] B. Benichou, T. Gils and K. Vranckaert, Design thinking in the legislative process: the key to useable legislation?, April 2021, https://www.law.kuleuven.be/citip/blog/design-thinking-in-the-legislative-process/. See also: T. Gils, K. Vranckaert and B. Benichou, "Exploring Policy Prototyping – Some Initial Remarks", July 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3885571

[2] European Commission, Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 21 April 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

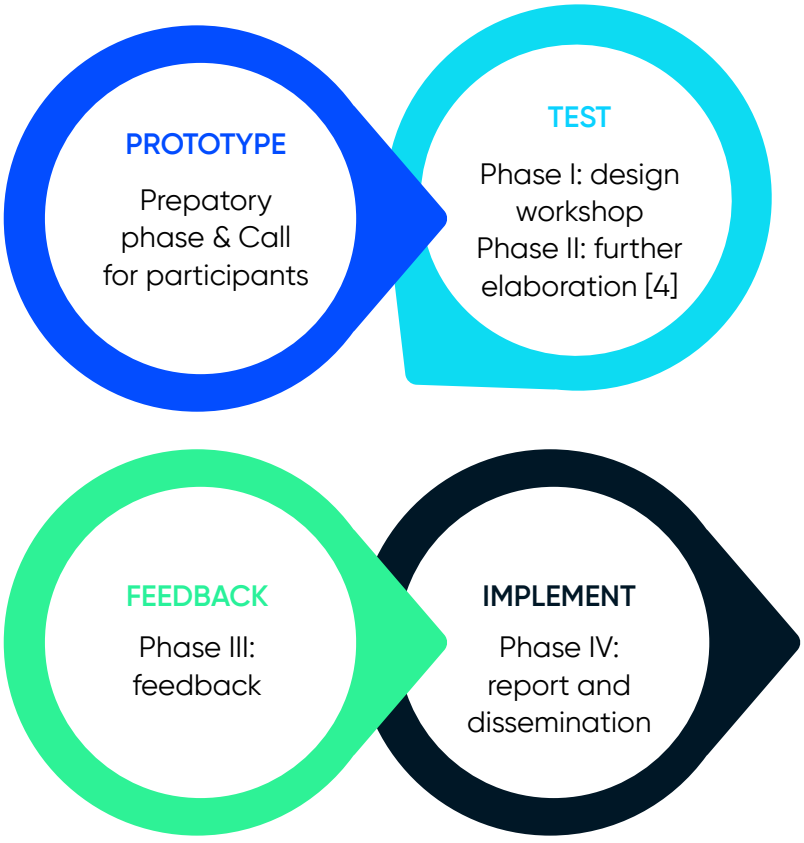
[3] URL: https://data-en-maatschappij.ai/en/news/policy-prototyping-an-assessment-of-articles-5-6-of-the-eu-ai-act

4. POLICY PROTOTYPING: METHODOLOGY AND PROCESS

In this part we will explain the methodology that we followed for this policy prototyping project. We believe that this is necessary to enable a correct interpretation and use of the results included in this report.

The policy prototyping project outlined in this report was initiated in the spring of 2023, commencing with an initial phase dedicated to the selection of the 'policy prototype' to be tested (i.e. the AI Act). Subsequently, a call for participants was issued, and interested stakeholders were identified. This group was invited to a design workshop, during which participants collaborated in small groups to draft mock compliance documents (based on real use cases) that would be required under the EU AI Act. These documents were further elaborated over the summer of 2023. Both the policy prototype as well as the prototype compliance documents were then subject to feedback via a qualitative online survey or in-person interviews. The findings of that survey are aggregated in this report.

The visual below illustrates how our phases map on the (theoretical) phases mentioned in part 3.1.



[4] The design workshop and the further elaboration could also feature in the prototyping phase as we created prototype compliance documents in those phases. However, as our main goal was to test the AI Act's requirements, we decided that they rather fit under the testing phase.

4.1. Preparatory phase: decision on legislative framework and practical considerations

The EU emphasizes transparency as a fundamental value for the development, deployment, and utilization of AI systems. The topic was already given prominence in the policy documents that preceded the AI act, such as the Ethics Guidelines for Trustworthy AI, issued by the High-Level Expert Group on AI, or the White Paper on AI, issued by the European Commission [5]. The AI Act continued that line. For this reason, we decided to focus this policy prototyping project on the transparency obligations in the EU AI Act proposal. More specifically, this project concerned the transparency requirements for high-risk AI systems (art. 13 AI Act) and the transparency requirements for "certain AI systems", including interactive AI systems, emotion recognition and biometric categorization and AI-generated/deep fake content (art. 52 AI Act). As we started this exercise, only the Commission proposal (April 2021) and the Council-text (December 2022) were available. The European Parliament-text (June 2023) was published when this exercise was already ongoing. Therefore, we decided not to include its amendments to these articles in this project [6].

Furthermore, we had to consider several budgetary and practical considerations. Although we welcomed international participants, the Knowledge Centre does not have the financial capacity to cover international travel expenses. We reimbursed local travel expenses of participants and allowed international participants (who could not travel to Belgium) to contribute virtually during phase III (feedback). In that way, we tried to lower the threshold and enable European and international participation. Apart from the local travel cost reimbursement, we relied on the voluntary commitment of participants and did not pay anyone. Since we relied on such a voluntary commitment, we deemed it appropriate to provide participants with an estimated effort at the start of the process. We will elaborate on this in the following section.

4.2. Call for participants

In order to try to ensure a diverse and representative group of participants to our exercise, we combined a public call for participants with targeted invitations to organisations or actors that we believed would or should be interested in our project.

[5] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 8 April 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
European Commission, White Paper on Artificial Intelligence, 19 February 2020, https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[6] The same applies for all the later compromise texts or leaks that were published or circulated. Important to consider, is that these later texts often did not contain significant amendments to the discussed articles.

The public call for participants contained the following information:

- On the one hand we looked for **"interested stakeholders/parties"**, incl. companies using or developing AI, (end) users, civil society, advisors,…. This type of participants was expected to primarily function as a test panel and sounding board. For instance, we aimed to offer AI-providers the possibility to test their AI application against the prototype(s) that would be developed during this project, while end-users and other stakeholders could assess if the information provided suffices their needs.
- On the other hand, we looked for **"experts"**, which we understood as experts with (practical) experience/expertise in facilitating transparency in a technological context. Their primary function was to co-create and develop the prototype compliance documents. Through participation, we aimed to provide them with the opportunity to engage with interested stakeholders and improve their skills.

We expected the efforts of participants to be different depending upon whether they were a (end-) user/provider of high-risk AI systems or an expert. In terms of time investment, we estimated that experts would spend about 2 to 3 working days in total (attendance design workshop, further elaboration of prototypes and intervention in the feedback phase III). Other participants would probably have been able to manage with a more limited time investment, as they were not expected to contribute to the further elaboration of the prototypes. In practice, however, these roles were not strictly applied and there were several groups that collectively further elaborated the prototypes.

4.3. Phase I: Design Workshop

As a first step, we organised a legal design workshop which was conducted purely in-person in order to ensure meaningful personal interaction. 17 participants and 4 facilitators worked together for a whole day in five different groups to shape first versions of different prototype compliance documents.

Every group focused on a single use case under either the transparency requirements of art. 13 AI Act (three groups) or art. 52 AI Act (two groups). These use cases were provided by the providers/developers of AI-technology involved in the exercise and based on their own, existing AI-offering. This ensured that the prototyping exercise had a sufficiently concrete angle and that we could really test during the workshop how feasible and practicable it is to include all the required information into understandable and clear instructions for use. We did not expect participants to release technical or sensitive details in relation to their use cases. It should be underlined that we looked for high-risk AI systems under the AI Act (for art. 13 AI Act) and AI systems that would be considered an interactive system (e.g. a chatbot), an emotion recognition system or a biometric categorisation system, or deep fake technology (for art. 52 AI Act). The use cases are explained in-depth below (part 6).

Three groups worked on **prototype Instructions for Use (IFU)** under art. 13 AI Act. Participants had to draft an IFU in accordance with art.13, §3 AI Act, while taking into account the requirements from art.13, §1 and §2 AI Act.

The other two groups worked on **prototype disclaimers** for AI systems falling under art. 52 AI Act. These groups also needed to come up with a related prototype process that allowed to decide when/whether or not to apply the disclaimer.

The workshop followed a legal design methodology, building further on our previous experience with legal design workshops [7]. In practice, this means that the workshop had four parts.

1. Empathize	The first part focused on understanding the technical use case and its environment. It also included mapping the affected stakeholders for every use case (incl. users) and their concerns.
2. Define	During the second phase, participants defined the problem that needed to be resolved. This included considering questions such as: what must be in the prototype? Which (legal or practical) requirements may be difficult to include? Are there aspects of the system's environment or users that are an issue for the prototype? When will the prototype be used?
3. Ideation	The ideation phase served to brainstorm about possible solutions to the problems defined in the previous phase, while taking into account e.g. the affected stakeholders and their concerns. At the end of this phase, possible solutions were clustered, prioritized and a choice was made regarding the prototype that would be developed.
4. Prototyping	During the last phase, participants started to work on an actual prototype. As participants knew that prototypes would be further developed during the next stage in the policy prototyping project, they focused on agreeing on the structure and substantive foundation of the prototype.

[7] Legal design experience: https://data-en-maatschappij.ai/en/event/workshop-legal-desgin

A total of 21 persons participated in the workshop: 4 facilitators from the Knowledge Centre, 8 representatives of AI-developers (incl. start-ups) and -users, 2 consultants, 4 legal experts, 2 representatives of civil society organisations and 1 academic researcher. Other people also showed their interest but could not make it to the workshop. They were invited to contribute to the later feedback phase.

4.4. Phase II: Further elaboration of prototype compliance documents

The design workshop was followed by a second phase, during which the prototype IFUs and disclaimers created during the workshop were further developed by the respective team members. This phase took place over the summer of 2023. While doing so, they closely took into account the requirements in the draft AI Act, as used during the workshop. The idea of this phase was to create well-developed, but not necessarily final prototype compliance documents on which comprehensive feedback could be given.

4.5. Phase III: Feedback phase

Once the prototype IFU and disclaimers were delivered, we launched phase III of the policy prototyping project: the feedback phase. In order to further diversify potential feedback, we decided to publish a second call for participants. This call did not distinguish between types of participants and aimed at attracting professionals and experts in AI. In practice, we primarily attracted additional legal professionals or service providers. People who signalled their interest to participate during the first call for participants but could not attend the design workshop were also invited to contribute to this phase.

We gathered feedback on both the created prototype policies as well as the related legal requirements. Participants were able to provide feedback (i) on how the prototypes implemented the requirements of the AI Act, and (ii) on the practicability, feasibility,... of the legal requirements themselves. With regard to this second aspect, we especially solicited feedback from participants who took part in earlier phases in order to capture their view on the implementation of the AI Act requirements into their own prototype.

Feedback was gathered in two ways:

- **Interviews:** participants had the possibility to opt for an in-person interview conducted by staff of the Knowledge Centre. Only a minority of participants chose this option.
- **Online survey:** the majority of participants opted to fill out the online survey. They were given a three-week timespan to complete the survey.

Participants were divided into two groups, with each group focusing on one article of the AI Act and the related type of prototypes. Every group had access to all the prototypes created under that article (i.e. art. 13 AI Act – 3 prototype IFUs; art. 52 AI Act – 2 prototype disclaimers and related policies), allowing them to compare the various prototypes.

Furthermore, they also received an extensive instruction e-mail and a briefing document. This briefing document contained some background information on policy prototyping and the current project, the description of the use cases and the text of the AI Act as used during the design workshop. We explicitly asked (new) participants to familiarise themselves thoroughly with the use cases, the related prototypes and the applicable legal requirements before starting the survey.

15 people contributed their feedback: 2 representatives of AI-developers (incl. start-ups) and -users, 4 consultants, 4 legal experts, 1 representative of a civil society organisation, 3 academic researchers with a technology background and 1 academic researcher with a legal background. This was a disappointing number as 37 people registered their interest to contribute feedback. This obliges us to acknowledge that the results and feedback, while interesting and valid, may not be highly representative or generalisable. We will evaluate and improve our feedback process in light of future policy prototyping exercises.

4.6. Phase IV: Report - publication of feedback and lessons learned

This report is the final stage of our policy prototyping project. It contains our findings, based on aggregated participant feedback, and lessons learned regarding the implementation of the transparency requirements. The intended audience of this report is (i) policymakers involved with the AI Act and its implementation, (ii) supervisory authorities that will be involved in the future enforcement of the AI Act, (iii) all stakeholders that will need to comply with, or benefit from these requirements and, (iv) all other interested parties.

This report is driven by multiple objectives. Primarily, it aims to assist stakeholders and professionals to effectively operationalise the transparency requirements of the AI Act by offering examples for Instructions For Use and Disclaimers, coupled with best practices and valuable lessons learned. Additionally, it seeks to convey the insights gathered from this project to policymakers and authorities, providing them with a practical perspective that could be instrumental in improving the AI Act's future implementation. Lastly, the report contributes to the evolving conversation on policy prototyping, advocating for its significant value as a tool in the policy development process.

5. THE AI ACT – TEXT OF ART. 13 & 52 AI ACT

Below we include the AI Act's text as it was used by the participants. *Green highlight* indicates that these parts were added by the Council, in comparison to the Commission's proposal. Blue highlight, on the other hand, indicates parts in the Commission-text that were removed by the Council.

5.1. Article 13 AI Act (European Commission and Council)

- **§1 – General principle:**
  - High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the output *(/system)* and use the output *(/system)* appropriately, and (appropriate type & degree of transparency will be ensured) to achieve compliance with obligations of user and provider
- **§2 – Obligation to draft instructions for use:**
  - High-risk AI systems must be accompanied by instruction for use. These Instructions must be:
    - in an appropriate digital format or otherwise;
    - include concise, complete, correct and clear information;
    - that is relevant, accessible and comprehensible to users.
- **§3 – Mandatory information**
  The information in the instructions of use must include:
- identity and contact details of the provider of the high-risk AI system
- the characteristics, capabilities and limitations of performance of the high-risk AI system, including:
  - (i) intended purpose, *inclusive of the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used;*
  - (ii) level of accuracy *(including its metrics)*, robustness and cybersecurity against which the system has been tested and validated and which can be expected, any known and foreseeable circumstances that may impact these expected levels of accuracy, robustness and cybersecurity;
  - (iii) any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights;
  - (iv) performance regarding the persons or groups on which the system is intended to be used; / *when appropriate, its behaviour regarding specific persons or groups of persons on which the system is intended to be used;*
  - (v) when appropriate, specifications for the input data, or other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose
  - *(vi) when appropriate, description of the expected output of the system*
- changes to the system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;

- human oversight measures, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users;
- *the computational and hardware resources needed*, expected lifetime of the high-risk AI system and any necessary maintenance and care measures, *including their frequency*, to ensure the proper functioning of that AI system, including as regards software updates.
- *a description of the mechanism included within the AI system that allows users to properly collect, store and interpret the logs, where relevant*

5.2. Article 52 AI Act (European Commission and Council)

Certain AI systems must meet specific transparency obligations under the AI Act.

- **§1 - AI system intended to interact with natural persons**
  - AI systems intended to interact with natural persons must be
    - designed and developed in such a way that natural persons are informed that they are interacting with an AI system
    - unless this is obvious from *(the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account)* the circumstances and context of use
  - This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, *subject to appropriate safeguards for the rights and freedoms of third parties*, unless those systems are available for the public to report a criminal offence.
- **§2 - Emotion recognition systems or biometric categorisation systems**
  - Users must inform exposed natural persons of the operation of the system
  - This obligation shall not apply to AI systems used for biometric categorization/ *emotion recognition*, which are permitted by law to detect, prevent and investigate criminal offences, *subject to appropriate safeguards for the rights and freedoms of third parties*
- **§3 - AI systems that generate or manipulate images, audio or video content that appreciably resembles existing persons, object, places or entities or events, and would falsely appear to a person to be authentic or truthful (deep fakes)**
  - User must disclose that content was artificially generated or manipulated;
  - Unless where the use:
    - is authorised by law to detect, prevent, investigate and prosecute criminal offences, or
    - is necessary for right of freedom of expression and right to freedom of arts and sciences + and subject to appropriate safeguards for the rights and freedoms of third parties / *where the content is part of an evidently creative, satirical, artistic or fictional work or programme subject to appropriate safeguards for the rights and freedoms of third parties.*
- *§3a – Requirements*
  - *information referred to in paragraphs 1 to 3 shall be provided to natural persons*
    - *in a clear and distinguishable manner*
    - *at the latest at the time of the first interaction or exposure*

6. RESULTS

6.1. Introduction

In the first phase of the project, we invited stakeholders (incl. AI-developers and (legal) experts in policy drafting) to a legal design workshop. During this workshop, the participants started creating the prototype compliance documents. Five AI-providers outlined their use cases and engaged with their group of experts to clarify the transparency requirements in the AI Act and their practical implementation. These five groups then further elaborated the prototype policies for their respective use cases in the second phase of the project.

- Three groups (Group I) have worked on prototype **Instructions for Use** (IFU) under art. 13 AI Act. It always concerned a specific high-risk AI system under the AI Act for which participants drafted an IFU in accordance with art.13, §3 AI Act, while taking into account the requirements from art.13, §1 and §2 AI Act.
- The other two groups (Group II) have worked on prototype **disclaimers** for AI systems that would fall under the transparency requirements of art. 52 AI Act. They also needed to come up with a related prototype process that allows to decide when/whether or not to apply the disclaimer.

As explained above, we both collected feedback regarding the individual prototypes as well as regarding the regulatory requirements of the AI Act. In parts 6.2 and 6.3, we will elaborate on the feedback on the individual prototypes and identify respective best practices and other lessons learned for each type of prototype. In part 7, we will then discuss the substantive feedback on the AI Act.

The prototype compliance documents themselves can be found in the separate annex to this report. We suggest to keep them close in order to facilitate the reading process, as the feedback often refers to specific parts of the respective documents.

6.2. Prototype Instructions for Use

In this part we will present the three IFUs that were drafted in accordance with the AI Act's requirements (see above). Each IFU relates to a separate use case but takes into account the same requirements. The IFUs are highly text-based so we suggest reading through them at least superficially before turning to the respective feedback.

6.2.1. Instructions for Use 1 - AI software for detection of eye pathologies

Use Case - AI software for detection of eye pathologies based on eye retina images

This use case focuses on AI software that analyses eye retina images in order to assist ophthalmologists and healthcare teams in identifying and/or diagnosing certain specific diseases, such as diabetes-related eye pathologies. It entails the usage of retina cameras, uploading the images to a web portal, the performance of the analysis of the images by the AI software and the interpretation of results. The IFU further explains the use case and is primarily targeted to ophthalmologists ('eye doctors') and other medical staff.

This prototype IFU is generally considered user-friendly and informative (8 out of 9 reviewers) and a majority believes it succeeds in creating transparency regarding the use case (6 out of 9). More specifically, reviewers consider this IFU to be comprehensible and accessible. Specific elements that receive positive feedback in this regard are: (i) the structure of the document and ranking of the topics; and (ii) the fact that certain paragraphs focus on separate users (e.g. parts 5.1.4-6 IFU), improving the accessibility of the document. Although basic and textual, the design of the IFU did not attract negative comments, which could confirm that the content is sufficiently accessible, not requiring visual aids.

Reviewers believe that this IFU manages to strike a good balance between providing enough information and avoiding information overload. The prototype is generally found to be clear and concise by most respondents and they also agreed that the IFU is clearly targeted at the medical professionals that would be using this type of AI system (i.e. ophthalmologists as primary users). However, even for them, terms like 'SaaS' (p.2 IFU) or 'broadband' and 'supporting docker (p. 4 IFU) may be too technical, as mentioned by one reviewer. Another reviewer did not think that it was sufficiently clear who the primary user of this AI system is, referring to p. 4 IFU where both 'healthcare provider' and 'ophthalmologist' are mentioned [8]. Another element that received praise by reviewers, is the presence of a troubleshooting section (part 8 IFU) and the reference to a separate, additional 'manual' (p. 4 IFU).

At the same time, other elements were considered to be less clear and would merit attention or clarification. One reviewer warns about the rather repetitive and potentially confusing wording regarding the intended purpose (see the product description and intended purpose in part 3 IFU and the clinical benefits in part 4 IFU). Relatedly, this reviewer raised the question for which diseases this system could actually be used [9].

A majority of reviewers (6 out of 9) believe that the IFU complies with the requirements from art. 13 AI Act, while the three remaining reviewers indicate that it only partially

[8] It can be pointed out that part 3 IFU explicitly indicates ophthalmologists as the primary users while healthcare providers are considered secondary users.

[9] Currently, the IFU only specifically mentions 'diabetes-related eye pathologies' but it leaves room for interpretation.

complies with said requirements. Positive feedback was awarded to (i) the part on human oversight (part. 6 IFU) and its explicit distinction between organisational and technical measures; (ii) the explicit mention of retention periods in part 5.1.8 IFU [10]; and (iii) the explicit mention of the accuracy metrics (sensitivity and specificity) and information regarding the expected output (parts 5.1.2-3 IFU). Other reviewers suggest to further elaborate the level of robustness (part. 5.1.3 IFU) or wish to see more detailed accuracy metrics (e.g. per disease, especially if the AI system would be used to detect various diseases).

A first part that attracted critical comments is the section on input and training data. Several reviewers commented that part. 5.1.1 contained insufficient information regarding the input and training data, citing e.g. confusing wording concerning the respective sources of the input and training data and a more general lack of detailed information regarding the training data. Regarding the latter, reviewers seemed to have preferred to be able to read more information about collection modalities and data quality. Some reviewers also wished to read more information regarding personal data management [11].

A second part that drew criticism, is the part on cybersecurity (part 5.1.9 IFU). Several reviewers note that the description of the cybersecurity measures is too generic without a clear, thoughtful allocation of responsibilities, while they warn that the end user appears to be assigned too much responsibility [12].

Finally, reviewers pointed to two recurring under-reported elements (see below, part. 7.2.4). Firstly, several reviewers argue that in case the AI system at hand would be used to detect other diseases than diabetes related eye pathologies, that such change in scope should be considered a 'change' in accordance with art.13, §3 AI Act and should, therefore, be included in the IFU (which echoes previously mentioned comments in this regard). Secondly, one reviewer suggests explicating within which jurisdiction this IFU applies.

[10] Although this IFU deals with personal data (and one would therefore need to specify a retention period in e.g. a privacy statement in line with GDPR requirements), it can be noted that AI systems do not necessarily use personal data. The concerned reviewer therefore noted that also for non-personal data, it is often relevant to know how long those data would be stored as different retention periods may apply.

[11] Part 10.1 IFU refers to the privacy policy, but this reference was likely deemed too limited.

[12] One reviewer pointed out that multi-factor authentication would be a recommended practice, given the highly sensitive nature of the data processed (i.e. retina images) in this use case.

6.2.2. Instructions for Use 2 - Medical device for cardiac arrhythmias prediction

Use Case – Medical device for cardiac arrhythmias prediction

The use case focuses on an AI algorithm that can predict cardiac arrhythmias, irregular heartbeats that occur when the heart's rhythm is abnormal, potentially disrupting blood flow and causing symptoms ranging from palpitations to severe complications like stroke or heart failure. The AI application enables early detection and thus provides the potential for better follow-up of high-risk patients. The application will process the entered patient data (values of blood pressure and cholesterol) and the electrocardiogram (ECG) data from a self-testing application. It will then generate a prediction regarding the likelihood of atrial fibrillation. The prototype is oriented towards the user/deployer of the AI application, more precisely medical staff/practitioners.

The second prototype received the highest ratings from the reviewers. All reviewers indicated that the prototype is user-friendly (in terms of concept, structure, language, and design). Eight out of nine reviewers found it sufficiently informative, and in terms of transparency, seven out of nine reviewers gave their approval. The IFU differs from the other two prototypes by providing extensive information on accuracy, metrics and logs and a number of preliminary suggestions aimed at a user-friendly design of the IFU. Its comprehensive and detailed nature was appreciated, and a majority of reviewers considered the use case compliant with transparency requirements, but there were also many suggestions for additions or improvements. For example, questions were raised about the sufficiency of detail regarding training data.

In terms of design, positive feedback is received, with a few suggestions for potential improvement. Recommendations include incorporating performance and metric paragraphs as annexes and considering offering personalized accuracy metrics based on individual characteristics of a person which introduces the concept of a tailored user experience. This entails, according to one specific reviewer, envisioning a website feature where users can input specific personal traits, subsequently providing them with insights into how well the device performs for users with similar characteristics.

Given the complexity of the subject matter of the use case, one reviewer recommends adding a glossary of terms. A reviewer highlights that the explanation of output in part 3.2 IFU (risk categories) and part 6.3 IFU reveals a mixture of synonymous terms, namely 'risk,' 'likelihood,' and 'probability,' as well as 'intermediate' and 'moderate.' Ensuring consistency and clarity in the use of these terms will enhance the precision and understanding of the content.

A much-discussed item of use case 2 among reviewers was the 'Accuracy' section and the detailed information on performance metrics and evaluation metrics. That comprehensive approach is mostly appreciated, but some reviewers questioned its usability for the end user. Several reviewers cite a need for more accessible language and more context to the metrics in this section.

Completeness and correctness concerns related to IFU2 include vague cybersecurity descriptions and potential issues with responsibility allocation. There is also a recurring demand, specifically at IFU2, for more details about the testing dataset. Moreover, the current short description, spread over two sites (Age and Gender Inclusivity p.3 IFU & Dataset Characteristics p.3-4 IFU), would also contain a contradiction according to a reviewer (tested on a diverse population vs Caucasian and middle-aged)

Discussions on the target audience emphasize, in the context of the use case, the need to include more references to scientific studies. This could help to prove the added value of the device in clinical trials, particularly in clinical trials. Another reviewer suggested that the use case could benefit from a more detailed description of the intended user.

In its entirety, IFU2 presents an engaging and versatile template. While some imperfections persist, the template substantially aligns with the specified requirements. With consideration given to any supplemental information offered, this prototype could be a commendable example for an Instructions for Use (IFU) document within the medical domain.

6.2.3. Instructions for Use 3 - HR talent matching tool

Use Case – HR talent matching tool

In this use case, a four-way talent matching API is used to connect job seekers with vacancies relevant to them. The AI system takes several criteria into account such as work experience, travel time, acquired skills,…. Information on these criteria is extracted by an embedder from the CV of candidates. The prototype IFU is focused on users/deployers of the AI system that match job seekers with job offerings of employers.

Feedback on the IFU for Use Case 3 indicated that almost all reviewers found the prototype to be user-friendly (8 out of 9 reviewers) although only a slight minority found it to be sufficiently informative (4 out of 9 reviewers). A majority (6 out of 9 reviewers) still found it to create sufficient transparency on the use case.

Participants indicated several elements which could have been added or elaborated to improve the comprehensibility and clarity of the use case. Specifically, participants suggested adding operating instructions and information on the interpretation of results in the IFU. More information on how the model was trained, as well as on the source of the data used, was also requested. The intended purpose could have been clearer. The robustness analysis of the use case was considered short by a participant. Further, participants indicated that the IFU's clarity is decreased by addressing both the user and jobseeker. Finally, the large amount of text in the IFU reduced its clarity.

Participants noted multiple times that the inclusion of an FAQ in IFU 3 improves its user-friendliness. The human oversight measures outlined in the IFU were considered clear by some participants but unclear by others. The inclusion of worst-case scenarios was considered a benefit by some participants (interpreted also as "foreseeable

circumstances" in conjunction with the section on human oversight by some participants) but considered strange, confusing, or unnecessary by others.

Participants also reviewed the completeness and correctness of the IFU as required by the AI Act. Participants indicated that some information required by the AI Act was missing. This includes information on the training data and its source. Metrics for the accuracy of the AI system, and on how biases were tested and evaluated, could also have been more detailed. The sections on the keeping of logs was not sufficiently clear. The instructions for the user of the AI system were also considered lacking.

A recurring remark by participants on this IFU was also that it was aimed at both users of the system and persons/job seekers affected by the system. This caused confusion since this is not explicitly stated in the document and results in (mostly) high-level instructions. The instructions could have been improved by adding a clear section on who the intended user of the AI system was and who the intended reader of the instructions is. Additionally, respondents suggested that multiple versions of the IFU could be made to address different affected persons.

6.2.4. General feedback on IFUs and common lessons learned/best practices

During the review process, respondents not only provided specific feedback regarding the individual IFUs, but also collective feedback applicable to all three IFUs. This general feedback encompasses various areas for improvement. One recurring suggestion involves improving the IFUs by incorporating links to related (legal) documents, such as privacy policies, and integrating visual elements like images, tables, or walkthroughs with screenshots (where possible or desirable). Simultaneously, participants acknowledged that the type and format of text documents as produced during this project will generally be the acceptable, base line market practice.

In terms of comprehensibility and accessibility, concerns were raised about the (overly) generic and technical nature of information in certain sections. Key points include the need for enhanced context, the use of clear and simplified **language** (while still providing detailed information) and the translation of IFUs into the local language. It is acknowledged that professional proficiency in English is often overestimated, necessitating a more inclusive approach. A positive element was the use of a table of contents in all IFUs and the overall clear structure of the documents. In order to enhance accessibility, one reviewer suggested standardizing and harmonizing different sections of the IFUs. Finally, participants express a preference for the inclusion of an FAQ and/or troubleshooting section.

Another recurring comment addressed the need to ensure that an IFU is clearly targeted at the envisaged, primary professional user. Reviewers find it useful that an IFU explicitly describes who the **primary user** should or is expected to be. Should an AI system allow for multiple types of users to be involved, some reviewers suggested to explicitly indicate the target audience for separate parts of the IFU.

The incorporation of worst-case scenarios, a feature employed in IFU 3, drew mixed responses from reviewers. While some participants expressed appreciation for this inclusion, noting that it enhanced the contextualization of the IFU, others raised concerns. Criticism centred on the perception that framing worst-case scenarios was peculiar and potentially confusing, with two participants expressing this viewpoint. Additionally, one participant argued that this aspect was redundant given its partial coverage in the human oversight section. In summary, the integration of worst-case scenarios appears to be subject of discussion and cannot be deemed a best practice for the moment.
In general, every IFU reflected the applicable **legal requirements**. As noticed by two reviewers, however, a shared, under-reported element was the required information regarding "changes to the system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any." It could be that in none of the use cases the developers anticipated any changes, while at the same time it can also not be excluded that the IFUs indeed overlooked this element or doubted about the meaning of this requirement (see also part 8.1) [13]. Another recurring topic attracting positive and negative comments was the detail of metrics included in the IFUs. Overall, respondents seem to prefer extensive and explicit information regarding the concrete levels of accuracy (such as sensitivity and specificity), performance and other relevant/related metrics (e.g. evaluation or bias/fairness metrics). If necessary, additional background should be provided to contextualize these metrics and enable a correct interpretation of results, depending on the target audience. One reviewer highlighted that the IFUs should be more specific about the geographical and functional setting within which the respective AI systems are intended to be used.

Furthermore, suggestions were made multiple times to provide more detailed information on input, training, validation and testing data and their respective source, related **data** management practices, personal data protection and cybersecurity measures. More specifically, respondents asked for more details regarding, at least, (i) the data used/required and its relevance, (ii) the data source or other collection modalities and (iii) data quality characteristics.

A remarkable observation is that all IFUs contain **installation and/or operation** instructions (be it limited) and that this attracted favourable comments, although not strictly being required by art. 13 AI Act. This observation does not necessarily surprise as installation and operation instructions (can) address different elements of the mandatory information under art. 13 AI Act (such as the intended purpose, circumstances that may lead to health or safety risks, performance or human oversight). Interestingly, one reviewer argues that installation and operation instructions (or troubleshooting) should not be included in an IFU but rather provided as a link or reference. After all, such information will likely have other owners (on the provider-side), other recipients (on the client-side) and another lifecycle, than the information required by art. 13. This can be understood as an argument in favour of working with a separate technical manual (especially if installation or operation instructions are detailed and lengthy).

[13] This can be illustrated by IFU1, where reviewers pointed out that in case the AI system would be used to detect other diseases than diabetes related eye pathologies, that this should be considered a 'change'. However, art. 13 AI Act limits the scope of this requirement by adding that such change should have been 'pre-determined by the provider at the moment of the initial conformity assessment'.

Reviewers also gave several suggestions on information they would like to see added to IFUs, although not strictly required by art. 13 AI Act. Suggestions include:

- An IFU should explicitly state the **jurisdiction** within which it applies (and contain related local contact details).
- IFUs should explain why the respective AI system is deemed **high-risk** (under the AI Act) and elucidate the related implications within the framework of the AI Act.
- IFUs should incorporate the concept of "**responsible disclosure**". Responsible disclosure entails the establishment by a provider of (i) procedures for reporting issues, particularly those related to data or cyber-vulnerabilities, and (ii) a dedicated, secure communication channel, which can be integrated into support or contact information, via which users can report these types of issues.
- An IFU should mention its **publication date.**
- IFUs should provide information related to the date and outcome of a **conformity assessment** (if applicable).

At the same time, quite a few other reviewers think that the information as required by art. 13 AI Act is sufficient and that it should not mandate additional information to be included in IFUs. In the same vein, it should be observed that several respondents ask for more details and information while often at the same time acknowledging that the risk of information overload is already high. This seems to confirm that there is indeed a need for comprehensive transparency requirements, but that they will only achieve their purpose if they are applied in a thoughtful and effective manner. In summary, striking the right balance between providing too much and too little information seems very fact-specific and difficult to achieve (see also part 8.1.1).

Finally, **other recommendations** complemented the IFU content. For example, one participant underlined that users should be reminded of IFUs (and, more broadly, the respective AI Act requirements) and that dedicated training should be provided to ensure sound understanding of the document and correct use of the AI system. This can be especially important to avoid misuse of data.

Overview lessons learned/best practices for IFUs

General lessons learned/best practices

- **TARGET THE PRIMARY USERS**: ensure that an IFU is clearly targeted at the envisaged, primary professional users. Explicate or describe who the primary user should or is expected to be. If the AI system allows for multiple types of users to be involved, consider explicitly indicating the target audience for separate parts of the IFU.
- **STRUCTURE**: ensure that the IFU follows a logical structure so that the target audience can understand the document properly without having to jump back and forth between sections. Start from a logical table of contents.
- **LANGUAGE:** strive to use simple, concrete and clear language, adapted to the target audience. Avoid overly generic wording or technical jargon (if possible). Consider drafting a glossary of terms if various, similar concepts need to be used. Provide translations to the local language.
- **DESIGN:** use a clear and legible font and page layout. Consider using visual elements such as images, tables or graphs (if possible and/or desirable) to enhance understanding.

Content-related lessons learned/best practices

- **SPECIFY METRICS:** provide extensive and explicit information regarding the concrete levels of accuracy (such as sensitivity and specificity), performance and other relevant/related metrics (e.g. evaluation or bias/fairness metrics). If necessary, additional background should be provided to contextualize the metrics and enable a correct interpretation of results
- **DATA:** ensure that information regarding input, training, validation and/or testing data is sufficiently detailed. Designate, at least, (i) the data used/required and its relevance, (ii) the data source or other collection modalities and (iii) data quality characteristics
- **INSTALLATION & OPERATION INSTRUCTIONS** / **TECHNICAL MANUAL:** include minimum installation and usage instructions in an IFU. If these instructions are very detailed or lengthy, consider using a separate, technical manual with the detailed installation and operation instructions while integrating a clear link to such document in the IFU.
- **HUMAN OVERSIGHT:** when discussing human oversight measures, consider making an explicit distinction between organisational and technical oversight measures.
- **CYBERSECURITY:** ensure that cybersecurity-related information is sufficiently concrete and actionable, following a realistic allocation of responsibilities.
- **CHANGES:** Be aware that an IFU should include information regarding possible future changes to the AI system and/or its performance.
- **FAQ/TROUBLESHOOTING:** consider adding a FAQ and/or Troubleshooting-section

6.3. Prototype Disclaimers and decision-making processes

In the following part, we will present two use cases falling under the scope of application of art. 52 AI Act and for which a prototype disclaimer and a related decision-making process were developed. The first use case relates to the use of chatbot and would have to comply with art. 52, §1 AI Act. The second use case involves the use of a deep fake and would have to comply with art. 52, §3 AI Act [14]. The respective requirements are similar, but not identical. This may be relevant when comparing the documents and feedback. Additionally, readers will observe a distinct divergence in approach between the prototypes for Article 13 and 52 AI Act. Notably, the prototype disclaimers discussed below adopt a more streamlined format enriched with visual elements.

As with the IFUs, we also surveyed the user-friendliness, informativeness and transparency level for the disclaimers. The answers to these questions were not conclusive and we have reasonable doubts whether some respondents recorded their answers correctly. Therefore, we decided not to include these findings in this report.

6.3.1. Disclaimer and decision-making process 1 - HR Chatbot

Use Case – HR Chatbot

This use case concerns a human resources (HR) AI chatbot intended primarily to communicate with employees (white-collar, blue-collar) or other people involved with a particular company (i.e. managers, outsourced staff) in relation to HR-related queries. The chatbot is trained on a collection of documents (e.g. internal HR documentation) and can answer questions related to this specific topic.

In the HR chatbot disclaimer, the authors limit information about the technology used while incorporating a visual representation of the requirements/user journey and the prototype decision-making process.

The first part of the prototype provides the user with an introduction to Article 52 AI Act. This is considered unnecessary, complex and unclear by several reviewers, indicating that a more practical approach is preferable to a theoretical, legal overview. In the general evaluation of comprehensibility and accessibility, two reviewers commend the decision to include a brief use context within the background information description (but would have liked it more elaborate). This not only enhances accessibility for readers of the prototype document as such but also aligns with the practical-focused nature of the second part of the disclaimer.

[14] We do not have a use case that would fall under the rules of art. 52, §2 AI Act (emotion recognition systems or biometric categorisation systems).

In that second part (the prototype user journey) a reviewer notes the absence of a clear explanation of how the disclaimer will be presented to chatbot users. Another reviewer wished to have seen a more elaborated disclaimer, not just the description of information that would be included, enabling a more substantive assessment (e.g. regarding the conciseness and clarity of the language used). Nonetheless, this brief description of information was deemed understandable. Furthermore, several reviewers argue that the disclaimer falls short in terms of being tailored to the target audience. According to one reviewer, this chatbot disclaimer is drafted so generally that it can be applied to any chatbot application or target audience (which would not necessarily be something negative in other circumstances) [15].

Additionally, two reviewers express their wish for receiving information on the technology used to build the chatbot (e.g. in the 'additional information' section) [16]. What is perceived positively, is the inclusion of a provision for complaints and feedback in the 'additional information' section. Also the flowchart approach is appreciated by one reviewer who adds that interactive elements could further enhance the design. Additionally, reviewers underscore the importance of testing the design of an avatar or interface with a diverse focus group to gauge (i) the effectiveness of the conveyed message, and (ii) its accessibility and digital inclusiveness (e.g. regarding blind people). In turn, such testing could lead to amending the user journey with possible reactions to the chatbot and accounting for different contexts of use and user abilities. Finally, a cautionary note from one reviewer emphasizes to refrain from humanizing AI tools, particularly chatbots (in line with prevalent calls from various organisations and academia).

Regarding the third part (the prototype decision-making process) one reviewer believed that the decision tree was concise, clear and well-balanced, while others believed it to be unclear, underdeveloped and not practicable, needing significant clarification.

Generally, respondents believed that this prototype provides a starting point to fulfil the relevant requirements of art. 52 AI Act.

[15] One respondent contemplated future improvements and envisaged the creation of "dynamically personalized/individualized yet privacy–preserving systems for contextual transparency notices, ensuring relevance without disruption". For instance, AI chatbots could adapt transparency disclosures based on their users' existing familiarity with chatbot interaction to avoid repetitive notices annoying experienced users.

[16] Such information could clarify if a chatbot is it built on a particular version of a large language model, or if it is a knowledge-based expert system.

6.3.2. Disclaimer and decision-making process 2 - Deep fakes in documentaries

Use Case – Deep fakes in documentaries

The use case involves the use of deep fake technology in documentaries. More specifically, the deep fake would be used to hide the identity of a victim testifying about domestic violence in television or online streaming documentaries. The use of deep fake technology (i.e. an AI-generated face) intends to protect the privacy and anonymity of people involved, while also retaining facial expressions and emotions. This should assist in preserving the authenticity of the message and facilitating audience identification, crucial for the prevention of domestic abuse. Given the potential size of the audience, the use case also had to take into account concerns related to vulnerable viewers, visually impaired, deaf or hard of hearing persons as well as persons with low literacy.

The second disclaimer takes a different approach and is more developed and detailed compared to the first disclaimer. This translates into the reviewers' assessment of the clarity, conciseness, accessibility and design of the prototype, which is unanimously positive. Logically, respondents consider that this prototype would fulfil the relevant legal requirements. Reviewers appreciated the use of the disclaimer's decision matrix, the first chapter of the prototype (AAA Matrix). By scoring various types of transparency tools in line with certain considerations (i.e. accessibility, anonymity and apprehension), the matrix clarifies which tools were preferred and further developed (i.e. an icon, information notices and a deep fake information policy) as part of the prototype disclaimer.

The following chapters of the prototype contain explanations and/or descriptions of the watermark icon, information notices and the deep fake information policy. This diversity of transparency tools, their concrete examples, the accessibility considerations regarding the icon (part 2 disclaimer), and the incorporation of different layers in the information policy (part 4 disclaimer) are explicitly lauded by the respondents. In the information policy text, only text marked as layer 1 would be visible at first. The text would expand after clicking on the title. In this way, the disclaimer enables the provision of detailed information while preventing information overload, according to a reviewer. This approach is deemed beneficial for tailoring the disclaimers' content to the target audience and avoid information overload, although one reviewer pointed out that the language used could be simplified [17]. Several reviewers also appreciated the detail on the type of technology that would be used to produce the deep fakes in the information policy.

At this point, it is interesting to note that one respondent, while supporting the approach taken by the prototype, exercises caution and suggests to study the impact of the multitude of transparency tools on the cognitive and affective aspects of the user experience. After all, this multitude of possible transparency tools could lead to information overload in its own turn.

[17] One reviewer suggests to "use explanatory multimedia to elucidate complex legal or technical concepts (e.g. using brief videos or interactive visualizations alongside text descriptions of how deep fake technology manipulates media)".

Another point of criticism, according to one reviewer, is that the disclaimer lacks information about a complaint and feedback procedure. Information on how to file a complaint or how to deliver feedback to providers should be added somewhere to pursue completeness.

6.3.3. General feedback on disclaimers and common lessons learned/best practices

In the feedback on the disclaimers, one could discern several overarching comments. A first set of comments emphasizes the importance of making disclaimers **accessible** to a diverse audience, particularly considering individuals with disabilities. A reviewer hereby referred to the latest WCAG guidelines, in which it is suggested that disclaimer messages may require different modes of understanding, especially for users who are blind. Concretely speaking, this will mean that disclaimers should not only be provided in writing, but also through aural and visual means, resulting in a diversity of transparency tools.

A second aspect concerns the provision of an **appropriate amount of information**. A striking observation based on this whole prototyping process is that although art. 52 AI Act, strictly speaking, only requires that natural persons are notified about the interaction with an AI-powered chatbot (§1) or, respectively, that deep fake content was artificially generated or manipulated, both prototypes go (much) further. Stakeholders (incl. AI providers) generally appear willing to provide more information. In addition, feedback participants also appear to positively value this additional information and consider it an acceptable level of information (preferring not to receive less information). At the same time, some participants warn not to exaggerate with this as too much transparency could also overwhelm users. Similar to the IFUs, **proportionality and balance** seem to be key when trying to implement transparency measures. A best practice that flows from the use cases appears to be the use of a layered approach both in relation to the various transparency tools used (e.g. a concise disclaimer which redirects to more detailed information), as well as in the disclaimer-related policies themselves. Some reviewers even suggest exploring interactive elements to enhance the clarity of disclaimer messages.

This eventually leads to a third aspect, **the target audience**. Respondents clearly find it important that disclaimers are adapted to the intended or potential target audience. This is not surprising since that intended or potential target audience has a significant impact on both accessibility considerations, and the appropriate level of information provision, and vice versa.

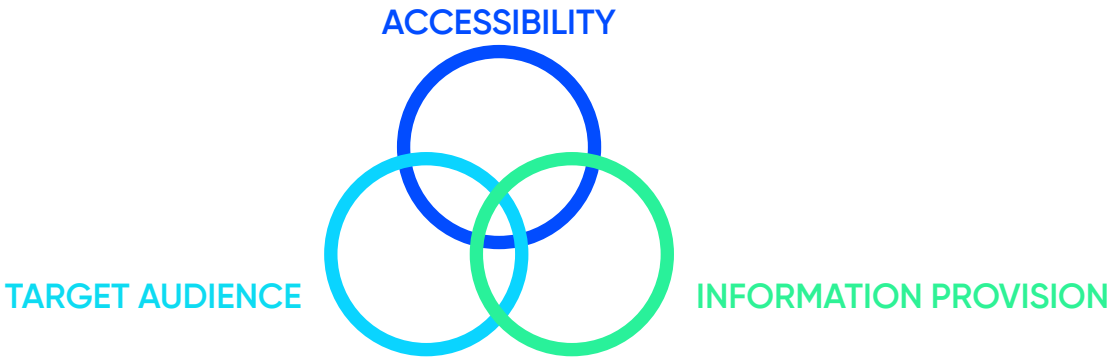


Figure 1-Visual representation of the reciprocal relationship between accessibility, information provision and target audience

Continuing on the theme of information provision, several additional elements or proposals were suggested by participants. For instance, they seem to favour that disclaimer-related policies include a section on how to file complaints or provide feedback to providers, additional details regarding the AI-technology used to create the chatbot or deep fake content [18] and a section containing a list of relevant references and applicable regulation.

In relation to the **prototype decision-making processes**, the collected feedback evidences that respondents clearly value the use of a visual and clear flowchart, decision tree or matrix, especially if these are able to take into account the variety of considerations and perspectives identified above.

A shortcoming of both prototypes is that they contain little explicit information regarding the impact of the art 52 **exceptions** on their decision-making process [19]. The chatbot use case does mention the respective exceptions but does not explain the concrete considerations they applied in their case (although they must have decided that the artificial nature of the chatbot would not be obvious). The deep fake use case does not mention the exceptions at all. This could imply that they considered it obvious or logical that they did not fall under the respective exceptions (for which there is a prima facie argument), but cannot be said with certainty. Moreover, one could integrate this into the AAA matrix as a preliminary consideration.

Overall, it is acknowledged that the prototypes demonstrate **promising transparency approaches** to the requirements of art. 52 AI Act. They could benefit from future elaboration with additional interactive media, individualization based on user needs around AI transparency and the further fine-tuning of the decision-making processes.

[18] We interpret the respective feedback to mean that such mention should be at least a concise and basic mention of the type of AI technique/tool used (e.g. the large language model used).

[19] Both §1 and §3 of art. 52 AI Act contain exceptions that relieve the provider of its transparency obligation (see part 6.2)

Overview lessons learned/best practices for disclaimers

General lessons learned/best practices

- **PROPORTIONALITY:** establish the desired level of transparency, taking into account the target audience and accessibility considerations. Develop and deploy a decision process that allows to identify which transparency tools are likely to achieve this objective in a timely, effective and proportionate manner. Consider that users seem to appreciate more than the strictly legally required information, but be cautious of providing too much information through too many channels.
- **ACCESSIBILITY:** Make sure the disclaimer is accessible to different users, taking into account users with disabilities. Follow established best practices and guidelines in this regard. This will probably imply that at least the disclaimer will have to be displayed in different forms (e.g. written, aural and (audio)visual)
- **TARGET AUDIENCE:** identify not only the intended but also the broader potential target audience to whom the disclaimer (and related information notices) may be presented. Avoid addressing the public in general and consider testing the design of the disclaimer (e.g. the avatar, icon or interface) with a focus group.
- **LANGUAGE:** Try to use clear, plain and concise language when drafting the disclaimer and the related information notices. Keep the target audience in mind when drafting the document.

Content-related lessons learned/best practices

- **USE A LAYERED APPROACH:** To optimize user engagement and avoid information overload, employing a layered approach in (information notices related to) the disclaimers might be effective. This method involves initially presenting only the essential information (layer 1) at first glance, with the option for users to expand the text for more detailed content (e.g. by clicking the title). This approach can be applied both in relation to the various transparency tools used (e.g. a concise disclaimer (layer 1) redirecting to a more detailed information notice (layer 2), as well as in the disclaimer-related notices themselves.
- **ALLOW COMPLAINTS/FEEDBACK:** consider including a provision for complaints and feedback in the disclaimer or related information notices. Facilitating complaints can be instrumental for improving transparency, fostering user trust, and ameliorating the disclaimer as this enables users to report issues, share their experiences, and suggest improvements.
- **SPECIFY THE USED AI-TECHNOLOGY:** consider providing concise and basic information in the information notices about the technology used to build the chatbot or create the deep fake content.
- **VISUALIZE THE DECISION-MAKING PROCESS:** develop a visual flowchart, decision tree or decision matrix that enables a deliberate decision regarding how and when to present a disclaimer or related information notices.

7. FEEDBACK ON AI ACT

In the following two parts, we will address and discuss the legal requirements of art. 13 and 52 AI Act in detail. During the feedback phase, respondents had the possibility to provide input regarding the practicability, feasibility, desirability and understandability of the requirements. We understand these concepts as follows:

Practicability

assesses whether the proposed requirements are realistic and can be implemented effectively in real-life situations without excessive difficulty.

Feasibility

involves evaluating whether the proposed requirements can be operationalised taking into account available resources and constraints, such as budget, time, technology, and manpower.

Desirability

evaluates whether the envisaged requirements and their operationalisation (i.e.. the prototypes) are useful and valuable to its intended audience or users.

Understandability

assesses how well the proposed requirements can be understood by the intended audience.

As feedback regarding practicability and feasibility often overlapped, we decided to merge these aspects and report on them jointly.

7.1. Feedback on Art. 13 AI Act

7.1.1. Practicability and feasibility

When asked if the transparency requirements included in article 13 AI Act where practicable, we received a variety of responses. Feedback on the practicability of the requirements ranged from positive (very practicable, many options to implement) to difficult to implement with major hurdles for AI-developers. Respondents were almost evenly divided between finding the requirements practicable (three respondents), somewhat practicable (three respondents) and difficult to implement (four respondents). One respondent abstained from commenting on the practicability of the requirements since they are not yet fully defined.

Respondents identified several specific implementation challenges relating to the transparency obligations. Prominent among the identified challenges is the **lack of concrete guidance**. Respondents refer to uncertainty about how detailed the information needs to be and the difficult balance between providing too much or too little information in the IFUs, ensuring they are both concise and complete. They highlight the uncertainty regarding the specifications of input and training data, and the cybersecurity, accuracy and robustness levels (including the respective metrics and the known and foreseeable circumstances that may impact such levels). Respondents also identified challenges in determining how "metrics" should be included. Specifically, it is not clear if providers should only include notice that "certain metrics were used" in the IFU or if they must also identify, justify and explain the specific metrics used. This uncertainty is further driven by the fact that balance must also be sought between different stakeholders both when drafting (i.e. a multidisciplinary team will have to decide which information to provide) and when targeting the documents at a particular audience (e.g. identifying the different knowledge levels and needs of stakeholders). Additionally, reviewers questioned how and when the instructions should be updated (e.g. if the intended purpose changes or evolves, does this automatically require updating? What about common software updates?) Reviewers therefore suggested to add further clarification to the article on when updates are needed. Finally, respondents also foresee challenges in providing relevant information on training, validation or testing data where this data may be sensitive.

Respondents were also asked to suggest solutions to improve the practicability of the transparency requirements of art. 13 AI Act. Many respondents suggested that additional guidance should be provided on how to interpret and implement the transparency requirements. This could take the form of:

1. Additional information in the **recitals**
2. Concrete **guidelines** by supervisory authorities
3. **Templates or examples** of IFUs (and other mandatory information) that are deemed compliant/best practice, possibly as the result of authority-endorsed policy prototyping initiatives
4. An official **assessment tool or automated review system** that provides feedback on draft instructions to providers. This could facilitate uniformity in the provided information in IFUs while allowing extraction by authorities of structured data for reporting and data driven policymaking.

Continuing this theme, respondents were also surveyed on the feasibility of the transparency requirements. Many respondents anticipate that the implementation of the requirements will create a significant workload (and financial burden) for the provider. To begin with, several respondents indicated that a technical background in AI is needed to be able to fully implement the requirements. This was confirmed by other respondents stating that implementation would require **a multidisciplinary team**, needing both technical experts and legal professionals. This may create issues for SME's or small start-ups, with implementation being difficult or impossible, due to a limited number of available financial and human resources. After all, enterprises that do not have specialised personnel will be required to hire and/or train staff or rely on third parties (e.g. consultants). For large organisations, the need for a multidisciplinary team can also be a

challenge as they will have to coordinate different departments within the organisation to focus efforts and knowledge to address the requirements. A minority of respondents considered the requirements more feasible, provided they are clearly communicated and guidance is provided. One respondent remarked that a company could and should already take the transparency requirements into account while developing and designing the system [20]. While additional efforts could still be needed to meet the requirements (e.g. drafting the IFU), such a **compliance by design** method would be beneficial for the provider. This was echoed by another respondent who stressed that providers will need to establish and use qualitative data collection and data management practices to proactively address the requirements.

7.1.2. Desirability

Most respondents found that the transparency requirements were **desirable**. This is motivated by the importance of informing users and affected persons about the system, creating trust in the system and forcing developers to reflect on important aspects of the development of the AI system. Several respondents noted that, while the requirements are desirable, they should be balanced to remain fair to the provider and to avoid overly technical explanations. Stakeholders implementing the requirements will otherwise have difficulties in balancing conciseness and understandability with the technicality and depth of the information to be provided. Some respondents also had reservations about the requirements, considering them less desirable in practice than in theory and only finding them desirable if they are part of an effective and clear legal framework.

One reviewer also pointed out that the increasing number of transparency requirements arising from distinct regulatory frameworks contribute to a general information overload. As a result, this makes achieving transparency an ever more difficult and complex objective. Furthermore, this reviewer also argued that this increases the risk of inaccurate or erroneous data creation which can lead to misinforming relevant authorities, markets and users.

Subsequently, respondents were asked if any requirements should be added to the transparency obligation in article 13 AI Act or if any of the requirements should be removed. A suggested addition was to require information about the **conformity assessment** that the AI system has undergone (e.g. when and by who) [21]. Another respondent suggested requiring transparency on how the user should act in case of (suspected) issues with the data or algorithm (incident response mechanisms) or, more generally, in case of user complaints [22]. Additionally, this respondent also suggested to require transparency on if, and how, input data would be reused in future **(re-)training and fine-tuning** of the AI system.

[20] This presupposes awareness and understanding of applicable requirements

[21] It should be noted that the transparency requirements are one of the requirements that should be assessed during the conformity assessment, resulting in a timing issue if this requirement was to be added.

[22]This reflects the identified best practice that IFUs should contain a FAQ-/troubleshooting section.

Another remarkable suggestion concerns requiring that IFUs should include detailed information on **the AI-technology used** (e.g. which algorithms, models or software) as well as on the applied explainability methods. Similarly, someone suggested to require transparency regarding the applied training and/or evaluation methods. It was argued that this could significantly enhance the value of IFUs and provide users with comprehensive insights into the functioning and decision processes of the AI system. A pointer regarding the acceptability and desirability of such requirement can be found in the prototype IFUs. After all, these documents demonstrate a willingness by providers and other stakeholders to disclose specific, technical information (such as detailed accuracy levels (e.g. sensitivity and specificity) and other metrics). It can therefore be wondered if transparency regarding the AI-technology used and applied explainability or training methods could and should not also be required by art. 13 AI Act [23]. However, it should be pointed out that while such information is not required by art. 13 AI Act (as discussed), art. 11 and annex 4 AI Act would already require (non-public, non-user oriented) technical documentation that contains this type of information [24]. A second, critical remark came from a respondent who argues that disclosing too much information on the AI-technology used (and data sets) may render an AI system more vulnerable to misuse, decreasing the desirability of any related requirements.

The following list contains all the suggested **additions to article 13 AI Act** by the feedback participants:

- Information about the conformity assessment that the AI system had undergone (e.g. when and by who);
- Information on how the user should act in the event of issues with the data or algorithm or in case of user complaints (including incident response mechanisms);
- Information regarding the (potential) (re-)use of input data for additional training/fine-tuning of the AI system;
- Information on the AI-technology used as well as on the applied training, evaluation or explainability methods;
- Description of a notice or procedure informing users that they can challenge AI-supported decisions and choose how to act (or not act) on them (e.g. as additional obligation related to the description of human oversight measures, as required by art. 13 AI Act);
- Mandatory references or links to the provider's other relevant policies (e.g. cybersecurity documentation or data protection-related documents).

[23] In order to be politically acceptable, such requirement would probably require a carve-out related to information covered by intellectual property or trade secret protection.

[24] Annex IV would require technical documentation to contain a detailed description of e.g. "the methods and steps performed for the development of the AI system, including, where relevant, recourse to pre-trained systems or tools provided by third parties and how these have been used, integrated or modified by the provider" and "the data requirements in terms of datasheets describing the training methodologies and techniques and the training data sets used, including a general description of these data sets, information about their provenance, scope and main characteristics; how the data was obtained and selected; labelling procedures (e.g. for supervised learning), data cleaning methodologies (e.g. outliers detection)" (both in the European Commission and Council text).

One respondent suggested to replace the requirement to provide information on the level of cybersecurity with a requirement to disclose which **recognized cybersecurity framework** providers implement.

Note that some respondents did not suggest changes or did not consider changes necessary. One respondent did not consider changes necessary now but was in favour of an iterative regulatory approach to renew the requirements if technological evolution requires this.

7.1.3. Understandability

Finally, we also asked respondents to provide feedback on the understandability of the language and terminology used in art. 13 AI Act. While multiple stakeholders indicated that they found the requirements understandable or accessible, most of these respondents also added that **AI-related technical knowledge** was a substantive benefit in understanding the requirements. The other respondents, a majority, found one to several points in the requirements which they did not find understandable. Respondents believed that many terms were not sufficiently distinguishable from each other and were too **interchangeable**. This included for example the differences between:

- "accuracy" and “performance”
- “robustness” and “cybersecurity”;
- “performance” of the system and its “behaviour”

In addition to confusion about specific terms, respondents also mentioned several points which require **clarification** in general. This includes, among others (as used in their respective paragraphs in article 13):

- “sufficiently transparent” (§1)
- "accessible and comprehensible" (§2)
- The description of the intended purpose of the system: what does “behavioural and functional setting” mean? (§3)
- "foreseeable circumstances"(§3)
- “reasonably foreseeable misuse”: how to understand “misuse” and when can it be “reasonably foreseen”? (§3)
- “risks to the health and safety or fundamental rights”: is this limited to risks for an individual or does this also encompass respective, societal risks (e.g. for public health or democracy)?
- "predetermined changes": what will constitute a “change” and when is it considered “predetermined”? (§3) [25]

Respondents also suggested clarifications regarding the part on AI system logs, such as how access to logs should be provided to users (i.e. whether this only requires that insights are provided or requires direct access to all logs) and whether such access should be free of charge or payment.

[25] As many AI systems will feature an important software-component, the obvious practical question is whether regular software updates should be considered as ‘predetermined changes’

With regards to the human oversight measures that should be included in an IFU, it was not clear for respondents whether the technical (or organisational) oversight measures must always include the possibility for overruling by a human.

To conclude, we highlight that respondents indicated that article 13 AI Act, and in particular paragraph 1 and the wording added by the Council, leave too much room for interpretation by the provider or were otherwise too broad and abstract. Furthermore, t**he relation between the paragraphs of art. 13 AI Act** was a cause for discussion: is the first paragraph, which establishes a general principle with general wording, a requirement in itself, or do the second and third paragraph further specify and limit the scope of the principle in the first paragraph to the listed elements? A question that will likely need to be resolved by supervisory administrative authorities or courts.

7.2. Feedback on Art. 52 AI Act

7.2.1. Practicability and feasibility

Feedback from respondents on the practicability and feasibility of article 52 AI Act was varied but in general we observe an attitude that the requirements are not insurmountable, nor straightforward to apply. Most respondents found them to be somewhat practicable, but the requirements raise several important concerns.

Participants believe the related workload to be manageable if providers are aware of applicable requirements, take those requirements into account early on and ensure the necessary commitment to implement them. An important, recurring concern in this regard is that adequate implementation of art. 52 AI Act will require **multidisciplinary** experience and knowledge across various domains. Participants stressed the need to have not only technical and legal professionals on board, but also professionals with experience in interaction design, user experience and accessible/inclusive design. These latter profiles are needed to ensure that information is conveyed in an engaging and accessible way to all possible recipients which, in turn, requires multimodal transparency (e.g. through aural and visual means). This is illustrated by a respondent who argues that e.g. a serious and officiously worded disclaimer may come across for vulnerable people as a legal protective measure by the provider, while it is important that the language should make clear that such transparency is part of an infrastructure of trust.

This leads to the second concern. As the wording of art. 52 AI Act is rather broad and allows for various interpretations, respondents foresee that the practicability of the requirements will be **use case-dependent**. In other words, the difficulty of implementing the transparency requirements will differ on a case-by-case basis, because different situations (and target audiences) call for e.g. different user experiences or imply distinct accessibility considerations. This is illustrated by art. 52, §3a AI Act which requires that information is provided in a 'clear and distinguishable' manner. Such a requirement is inherently audience-specific, so it may be difficult to think about all possible scenarios and anticipate all situations in abstracto.

Another aspect of this concern is that participants indicate that they find it difficult to identify the **appropriate amount of information** to provide to recipients, resulting from the broad and unspecific formulation of art. 52 AI Act. Participants indicate that they wish to properly inform users about the interaction with a chatbot or the artificial nature of a deep fake, but not overwhelm them while also avoiding completely disclosing their system's capabilities or characteristics [26]. In addition, it is pointed out that art. 52 AI Act gives quite an ample margin of discretion to chatbot and deep fake providers when deciding on implementation: do they prefer to provide only the strictly legally required information or do they prefer to do more? A respondent highlights that both use cases in this project go beyond the minimal requirements, but that they could have done otherwise [27].

When asked how the practicability of the requirements of art 52 AI Act can be improved, respondents suggested solutions similar to the ones for art. 13 AI Act. In general, they wish to have more (concrete) guidance and make multiple respective proposals:

- Additional information in the **recitals** and/or concrete **guidelines** by supervisory authorities.
    - These guidelines could take the form of a checklist or standardized criteria (ideally, use case-specific).
- Collaborative **pilot projects** with stakeholder involvement, aimed at identifying best practices and resolving unclarities (e.g. in the form of policy prototyping).

7.2.2. Desirability

Feedback participants unanimously stated that the transparency requirements of art. 52 AI Act are **highly desirable**. One respondent stressed that transparency is a cornerstone of 'responsible AI', implying that art. 52 AI Act rightly aims to promote trust in AI systems by requiring more transparency in relation to the applications in scope. Additionally, and also based on the prototype disclaimers, participants underline that it is desirable that providers provide more than the strictly legally required information. Especially as AI-literacy among people still varies greatly, it is imperative that that they are informed adequately about e.g. the interaction with a chatbot or the artificial nature of a deep fake image. Respondents otherwise fear that underprivileged groups or minorities will not be adequately aware of the technology they are exposed to.

Simultaneously, respondents provided a list of **possible amendments** to art. 52 AI Act. These suggestions are predominantly aimed at clarifying the wording and making it more fit for purpose:

[26] As evidenced by the prototype disclaimers, see part. 7.3.

[27] According to this respondent, possible minimal implementations are: (i) disclaimer 1 (chatbot) could mention that the chatbot is AI-powered only at the bottom of a page or via another screen only accessible via clicking "more info" after briefly informing that a chatbot is available; (ii) disclaimer 2 (deep fake) could inform the audience only at the beginning of the documentary that in some parts a deep fake will be used.

- One respondent suggests to merge the **exceptions** related to law enforcement/criminal investigation into one paragraph (e.g. merge the respective parts of §1,2 and 3 into a fourth paragraph). Another respondent explicitly acknowledged the added value of this exception for law enforcement/criminal investigation purposes.
- Another suggestion was to add a requirement that disclaimers should explain the **(un)expected output** of the AI system at hand (e.g. in order to frame possible concept drift). For example, a chatbot disclaimer would have to explicitly state on which topic(s) the chatbot can, and cannot, answer questions. A similar suggestion entails requiring that a disclaimer would feature a brief use case description which should explain the chatbot or deep fake's intended purpose.
- A third suggestion echoes the prototype feedback (see part 7.3.2) where various respondents asked for additional details regarding the **AI-technology used** to create the chatbot or deep fake content. In that context, one respondent suggests to legally require that disclaimers state the type of AI-technology/tool used for chatbots or deep fakes.
- A fourth proposal suggests to add "**accessible**" to art. 52, §3a AI Act (i.e. in a clear, distinguishable and accessible manner) in order to guarantee that accessibility considerations are taken into account when designing a disclaimer in order to achieve inclusive transparency.

7.2.3. Understandability

As apparent from the feedback regarding the practicability and desirability of art. 52 AI Act, the wording of the article is not entirely comprehensible and would benefit from some clarification. Interestingly, participant feedback seems to indicate that a **technical AI-background** is less of a prerequisite to understand and apply this article as opposed to art. 13 AI Act (see part 7.1.3).

Although respondents understand and acknowledge that the broad formulation of art. 52 AI Act allows it to cover a diverse set of technologies, some key concepts appear to be rather unclear for providers or other stakeholders to apply. Respondents highlight the following parts (as used in their respective paragraphs in article 52):

- "a natural person who is reasonably well-informed, observant and circumspect": how to determine this hypothetical figure when the target audience consists of a wide variety of people? (§1)
- How to understand 'subject to appropriate safeguards for the rights and freedoms of third parties'? (all paragraphs)
- "appreciably resembles": does this require the deep fake content to be identical to the source or is similarity sufficient? (§3)
- "falsely appear to a person to be authentic or truthful": does this presuppose an intent by the provider to delude or mislead the audience? (§3)

- "where the content is part of an evidently creative, satirical, artistic or fictional work or programme: how to apply this exception in the context of e.g. commercials using deep fakes of living or deceased celebrities? Arguably, a commercial can often be considered a creative or artistic work, while it would be good practice to inform consumers and other market actors about the deep fake nature of the celebrity representation. (§3)
- "clear and distinguishable": as already mentioned, participants highlight the relative and subjective nature of this requirement. (§3a)

Finally, one respondent argues that the reference to 'deep fake' in §3 may not be **future-proof** as we may use other expressions in the future.

7.3. General feedback

To conclude this part, we identify several common threads in the detailed feedback on the AI Act's transparency requirements.

It appears that respondents generally do not think that the requirements are entirely unfeasible or unpracticable. However, participants do underline that compliance will require **effort, proactive thinking and multidisciplinary experience and knowledge**. This latter factor will often make it less straightforward for smaller providers of the respective AI systems to deal with the requirements due to lack of available financial and human resources. It is therefore unsurprising that respondents ask for sufficiently concrete guidance, templates and examples to facilitate the implementation of the transparency requirements and level the playing field.

A large majority of participants believe that both sets of requirements (art. 13 and 52 AI Act) are **desirable** and support their inclusion in the regulatory framework. Transparency is seen as an essential tool to create trust. At the same time, participants provide many suggestions for amending both articles. A remarkable observation in that regard, is that participants wish to see additional requirements regarding the **disclosure of technical details** in both IFUs and disclaimers. Other respondents warn against requiring too much additional information as this may contribute to information overload and make AI systems more vulnerable. Finally, it is also argued that in order to maintain the desirability of this type of requirements that policymakers should keep legislation and policies up to date with technological changes.

Regarding the importance of understandability, we refer to the pertinent comment of two participants who independently stated that the value of the transparency requirements can be undermined by their insufficient understandability that creates **uncertainty** and a risk of misinterpretation, which in turn risks causing inconsistent application and/or superficial compliance. As evidenced by participant feedback on both art. 13 and 52 AI Act, both provisions contain many important concepts that require clarification.

8. CONCLUSION

As the EU advances towards a comprehensive legal framework for AI, the AI Act emerges as a pivotal regulation, seeking to balance the interests of innovation and society. The AI Act places significant emphasis on transparency requirements, mandating AI developers and deployers to disclose critical information about their AI systems. The idea behind this is that it will enhance accountability and public trust and empower regulatory bodies across EU member states to oversee the deployment and functioning of AI technologies.

This report offers comprehensive insights and practical guidance for policymakers, supervisory authorities, stakeholders and professionals who must work with the AI Act's transparency requirements' complexities. It has outlined five prototype compliance documents (three IFUs and two disclaimers) and presented critical stakeholder analysis and practical feedback. As demonstrated, each prototype comes with challenges and limitations, but they also allowed to identify some (early) best practices for complying with articles 13 and 52 AI Act.

Subsequently, the report provides detailed legal comments and feedback on articles 13 and 52 themselves, aiming to inform and guide policymakers and authorities.
In summarizing the feedback on Articles 13 and 52 of the EU AI Act's transparency requirements, several key points emerge. In general, participants confirm the desirability of the transparency requirements, acknowledging their role in fostering trust. However, they call for more concrete guidance and voice concerns about the potential for information overload and the need for legislation to evolve with technological advances. Furthermore, the understandability of these requirements was flagged as critical to prevent misinterpretation and ensure consistent application, underscoring the importance of clear and comprehensible legal provisions in AI regulation.

Throughout this project, the concept of policy prototyping has garnered positive feedback. Many participants recognized the significant value that policy prototyping may bring, thereby emphasizing the usefulness of exploring the application of regulatory requirements and provisions on an explicit fact-based use case. Participants agreed that this method can add much value to the policy implementation process. The positive reception underscores a broader consensus emphasizing the importance of actively involving a diverse array of stakeholders in the policymaking process. By gathering comprehensive insights from these varied perspectives, policymakers can ensure a solid foundation for the policy implementation process.

In the forthcoming months, the Knowledge Centre Data & Society is committed to maintaining a strong focus on the evolving topic of the AI Act through comprehensive analysis and focused events. By recognizing the significant future impact of this key legislative file, the KCDS considers it as a crucial area of interest. We will also continue refining and applying the policy prototyping method as a tool for innovation and foresight in policymaking. Additionally, the KCDS will organise a new policy prototyping project, with the specifics of the project set to be unveiled in the upcoming months. This initiative underscores our ambition to be at the forefront of policy development.

9. ACKNOWLEDGEMENTS / PARTICIPANTS

We would like to thank all participants whose contributions made the realization of this policy prototyping project possible. Your enthusiastic engagement and valuable insights were pivotal in shaping this project. Special thanks are extended to those who participated in the design workshop, investing time and expertise to draft mock documents. The collaborative efforts of everyone involved have been instrumental in the production of this report. Your commitment to advancing the discourse on policy prototyping in the field of AI and data policy is genuinely appreciated.

Project participants (1/2)

Benny Backx
IP/IT partner - Caluwaerts Uytterhoeven Advocaten

Marjon Blondeel
AI engineer – VUB AI Lab

Ellen Caen
Lawyer IP - Eubelius

Lynn D'eer
AI Project manager IDLab - imec

Nele Gerrits
Senior Researcher AI – imec

Lisa Koutsoviti Koumeri
PhD Researcher - UHasselt

Victor de Menezes
PhD. in Law - Universidade Federal de Santa Catarina (UFSC)

Anita Prinzie
Product Manager – Omina Technologies

Stef Rommes
CTO - Mona Health

Angeliki Tiligadi
Compliance & Data Protection Officer - Qover

Brahim Benichou
DPO – BW Legal BV / Delhaize

Elien Cadron
Cyber Security Professional

Elena Nunez Castellar
Assistant professor Intelligence Augmentation - Eindhoven University of Technology (TU/e)

Florent Diverchy
Marketing Intelligence Manager - Produpress

Dino Gliha
Partner – MGG Law

Mathias Leys
ML6

Luca Nannini
Minsait by Indra Sistemas, Madrid

Nele Roekens
Legal advisor AI & human rights - Unia

Maud Stiernet
Consultant - A Little Lining Comes

Wouter Travers
Senior Manager - PWC

Project participants (2/2)

Helen Tueni
Digital Transformation practitioner - Vucable

Joren Verspeurt
Machine Learning Engineer & Security Officer - Radix

Jan Yperman
R&D – VITO

Lisa Van der Aa
Project Manager – AML BV

Aagje Weyler
Compliance expert privacy & AI - imec

Knowledge Centre Data & Society Team

Thomas Gils
Research Associate – Centre for IT and IP Law/ KCDS

Wannes Ooms
Research Associate – Centre for IT and IP Law/ KCDS

Frederic Heymans
Research Associate – imec-SMIT, VUB/KCDS

Prof. dr. Jan De Bruyne
Co-director KCDS

POLICY PROTOTYPING

Knowledge Centre Data & Society