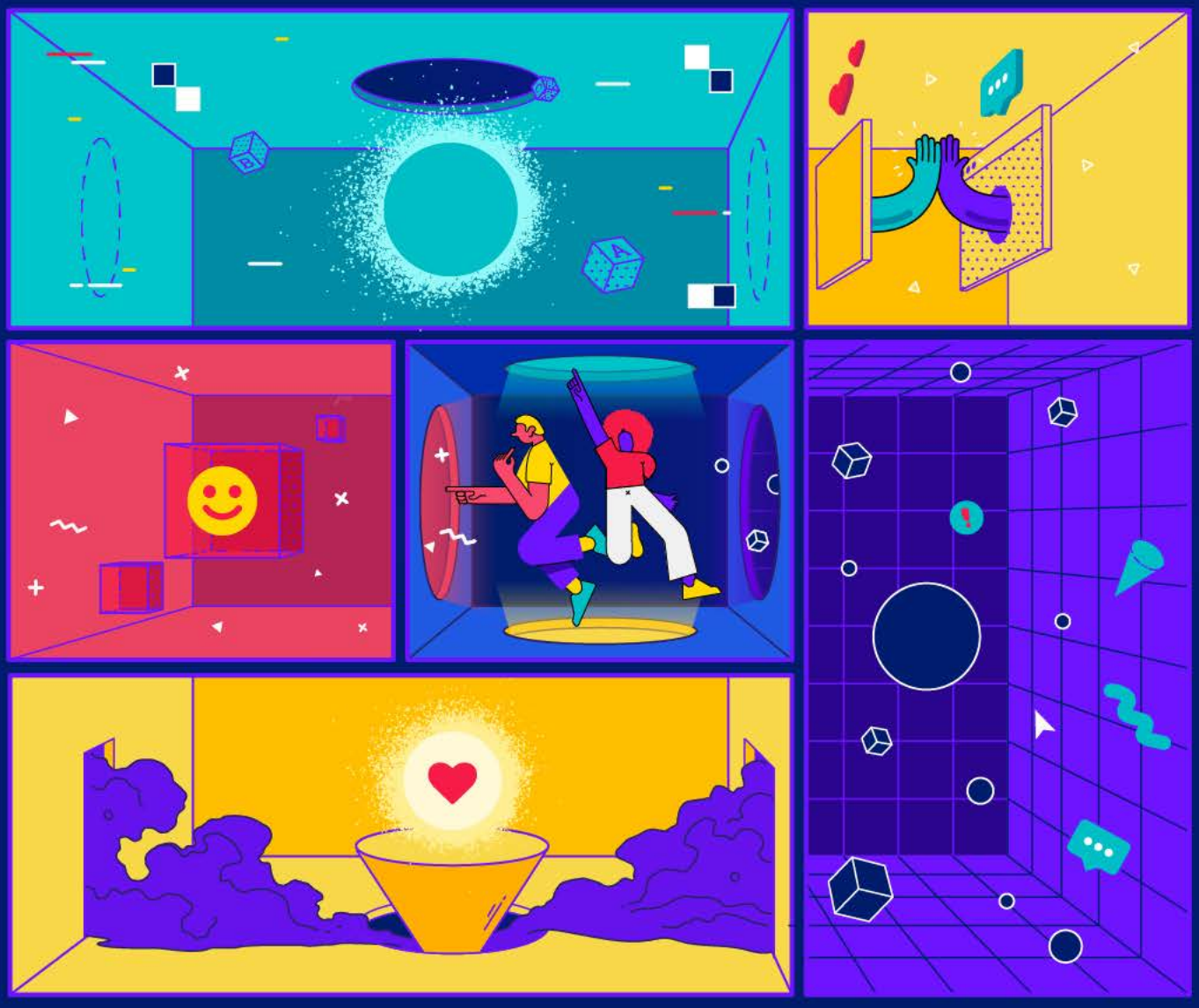
People-Centric Approaches to AI Explainability

ttclabs.net

Insights from product and policy prototyping with startups



This report presents the key findings from a series of co-creation workshops hosted by the Trust, Transparency and Control Labs (TTC Labs) and Open Loop in partnership with Meta Startup Programs and Singapore's Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC).

Meta TTC Labs OPEN LOOP INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Exploring practical approaches to AI explainability

Providing transparency and explainability of artificial intelligence (AI) presents complex challenges across industry and society, raising questions around how to build confidence and empower people in their use of digital products.

The purpose of this report is to contribute to cross-sector efforts to address these questions. It shares the key findings of a project conducted in 2021 between TTC Labs and Open Loop in collaboration with the Singapore Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC). Through this project TTC Labs and Open Loop have developed a series of operational insights to bring greater transparency to AI-powered products and develop related public policy proposals.

These learnings are intended both for policymakers and product makers – for those developing frameworks, principles and requirements at the government level and those building and evolving apps and websites driven by AI. By improving people's understanding of AI, we can foster more trustworthiness in digital services.

This report incorporates feedback provided by external experts during a consultation phase that took place over January and February 2022. If you have any further feedback or observations on this report, please reach out via ***info@ttclabs.net***

VERSION 2 • April 2022

What's in this report

Foreword

Global calls for greater transparency and explainability of AI systems are on the rise. At the heart of these discussions is the goal to foster public trust in AI through awareness and understanding of AI-enabled product outcomes. As international discourse on trustworthy AI progresses, countries and international organisations have started to publish principles and guidelines, and even to propose laws to govern the development and use of AI. Such a trend is pertinent to the AI developer community as it seeks to align with these soft and hard regulations.

The OECD's Principles on Artificial Intelligence are an important example of these developments as they articulate concrete recommendations for public policy and strategy, and can be signed up to by governments. Singapore, being an adherent to OECD AI Principles, has published a *Model AI Governance Framework (Model Framework)* which espouses that decisions made by, or with, the assistance of AI should be transparent and explainable, i.e. appropriate information is to be provided to individuals who may be impacted by the AI system. The *Model Framework* and its companion guide, the *Implementation and Self-Assessment Guide for Organisations (ISAGO)*, provide practical tools to help developers and users of AI systems do this responsibly.

A policy is only as good as its implementation. Similarly, the success of our *Model Framework* and *ISAGO* should be measured by the implementation experience of AI developers. As such, we are excited to be partnering with Meta's TTC Labs and Open Loop to translate key concepts of transparency and explainability into prototypes and apply them to real-world applications. What you see in this report are the fruits of the start-up programme between Meta and the Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC).

The example prototypes and solutions you will find in this report bring to life the concept of transparency and algorithmic explainability, particularly around stakeholder interaction and communication, a core focus of the *Model Framework* and *ISAGO*. From the perspective of user experience (UX) design, AI explainability has meaning for users of AI applications when it is executed as part of their experience, and the information is presented in context and in a manner that enables users to understand the process and take action where necessary. Applied well, it empowers users of AI applications and builds trust. It is by working with creative companies that we can start making constructive contributions to this field.

These interpretations will hopefully spur thinking on how to achieve the same within your solutions, and enhance accountability in the use of AI and data.

This journey has been a rewarding experience for all – from startups, to design mentors, to everyone else who has been part of this journey. We hope our practical, ground-up approach can be a positive contribution globally. We will continue to work together with industry in close collaborations to drive tangible results for the digital economy.

Mr. Yeong Zee Kin

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

Deputy Commissioner, *Personal Data Protection Commission*

Foreword

The current era of AI is marked by broad access to high-performance open-source tools and technologies. The route to testing and implementation is shorter than ever, and the ability for implementation of AI at scale by smaller companies is unprecedented. These innovations present enormous opportunities to develop groundbreaking products and services that are useful and meaningful to people in their everyday lives. These trends also present potential risks, as the necessary socio-technical approaches to making these technologies understandable, accountable and trustworthy are still in development. This creates a governance gap that requires collaboration from private industry, civil society, academia and government.

Since 2017, Meta's Trust, Transparency & Control (TTC) Labs has been co-creating to improve user experiences around personal data. TTC Labs brings together policymakers, privacy experts and product creators, using design thinking to improve trust, transparency and control in digital products. Driven by a user-centered approach, we seek to collaboratively prototype and test potential solutions that can work for people and for businesses.

Over these years, our team has led more than 50 co-creation sprints – known as Design Jams – around the world, collaborating with a wide range of stakeholders, reflecting diverse experiences, perspectives and disciplinary expertise. A highlight of this work has been our collaboration with the IMDA to work with startups across the Asia-Pacific to address key issues related to data transparency, as described in our 2019 report *People-centric Approaches to Notice, Consent, and Disclosure*. We are now building on that work with this report on algorithmic explainability, capturing insights from our program of co-design as well from leading research.

This process, and the outputs from these collaborations, are important for bridging the gap between these burgeoning technologies and the industry and government toolkit necessary to provide accountability, trust and understanding. Our collaboration with the IMDA has allowed us to go beyond conversation to practically explore usable approaches to these topics. This report shares our findings from co-creating with startups, through which Meta, via this work, has helped to test a product framework for AI explainability across sectors and at different scales, gathering valuable insights and practical experience. It's particularly critical that we explore and understand how to ensure compatibility between these frameworks and a useful, consistent and high-quality user experience – a key focus for our lab.

There is a common responsibility across large and small businesses to develop shared approaches to AI explainability. By working with a range of businesses at the forefront of the implementation of new AI technologies and services we can create a policy and product feedback loop that enhances the efficacy of both and leads to better outcomes for everyone who uses digital products.

Dr. Dan Hayden

Director of Data Strategy, *Meta*

Co-lead of *TTC Labs*

Introduction

Project overview

Investment in AI is soaring across sectors, together with increasing levels of academic research and government legislation on AI. What if a new wave of applied thinking around trustworthiness and transparency helped fuel responsible development for startups and established companies, both of whom are in the early stages of their AI technology and product development journeys?

The **People-Centric Approaches to AI Explainability** project brought startups from the Asia-Pacific region and Europe together with multidisciplinary experts in a series of co-creation workshops focused on product and policy prototyping.

These startups offer AI-powered products and services across a range of markets and sectors:

Betterhalf.ai (India)

A matrimony matchmaking app that uses AI to recommend personalized matches with minimal parental intervention

MyAlice (Singapore)

A customer support platform that uses AI to help e-commerce operators manage communications and sales across multiple apps and social media channels

The Newsroom (Portugal)

A news app that combats misinformation by using AI to curate a personalized newsfeed based on trustworthiness, objectivity and a person's specific interests

XOPA AI (Singapore)

A recruitment platform that uses AI and data science to remove bias and make hiring more equitable

Zupervise (UK, India)

A unified risk transparency platform to govern AI in the regulated enterprise

In Design Jam workshops facilitated by TTC Labs, the startups were joined by policymakers, privacy experts from civil society and academia, and other product makers to engage in collaborative product prototyping. The aim was for each multidisciplinary team to co-create a trustworthy AI experience – a design pattern aimed at improving people's understanding of the startup's AI and addressing any concerns they might have.

In policy prototyping workshops facilitated by Open Loop, these teams reconvened to test governance frameworks and to derive evidence-based inputs for future developments in policymaking processes.

Introducing the AI Explainability Framework

Meta's **Responsible AI team (RAI)** has been developing product design guidelines to better understand and take action on people-centric explainability experiences in AI-powered products and features.

The **AI Explainability Framework** featured in this report provides draft guidance for product makers on the design of explainability experiences. With principles and design guidelines structured around four dimensions of explainability, this Framework helps product makers keep the needs of people using or affected by AI-enabled products at the center of their design discussions. As a work in progress, the Framework is not reflective of practices at Meta.

For the **People-Centric Approaches to AI Explainability** project, the Framework was introduced to the Design Jam participants to use as a prompt, a reference and an analytical tool.

The prototypes co-created during these workshops show clear affinities with the Framework's structure and guidance. From a product design perspective, they variously make people aware that AI is involved in a digital product experience, explain individual AI-powered product outcomes, help people comprehend how a product works and provide further information about the underlying models.

The Framework has been central to the exploration of people-centric approaches to AI explainability over the course of this research. At the same time, this project has provided a practical and collaborative setting for validating the Framework, testing its guidance and principles across a range of use cases.

How we made this report

The aim of this report is to promote the adoption of trustworthy AI explainability practices. Testing and validating product and policy guidance, it provides a series of practical insights, considerations and observations derived from co-created explainability prototypes.

These prototypes – hypothetical design explorations using real industry products – were the primary research outputs of the Design Jam. Through analysis of these design explorations we generated the insights and considerations that form the basis of this report. Sharing these findings with a range of cross-sector experts, we revised and refined them in line with their feedback and suggestions.

The co-created prototypes are used throughout the report to illustrate key ideas and bring the insights to life. Alongside the personas developed by the participating startups, the hypothetical prototypes provide useful guidance on the potential implementation of the insights and considerations. They are supported by examples of real and fictional apps from previous TTC Labs Design Jams and Open Loop workshops.

Together these prototypes, personas and supporting examples highlight the role of design-led innovation for AI explainability, including the complex trade-offs and decision-making that occurs when designing for people's needs and real-life scenarios.

This open exploration of practical reasoning is intended to contribute to and advance industry and policy discussions about AI explainability.

Note

This report is not intended to be comprehensive. The Framework guidance and project insights contained within are neither reflective of current practice at Meta nor are they proposed as industry standard.

Getting clear on terminology

Trust	Sustainable relationships between digital products and the people who use them
AI Accountability	Mapping the person or entity responsible for each part of an AI system and to whom they are accountable
AI Governance	The set of policies and practices guiding the use of AI within an organization
AI Transparency	Being clear, open and honest about how an AI system is built, operates and functions through explainability and interpretability, ensuring outputs can be documented and scrutinized in order to hold AI systems to account
AI Explainability	Efforts to help people understand when and how they are engaging with AI systems
AI Control	Giving people meaningful agency over their relationship with an AI system

General Product Users

Anyone who uses a product or service as a consumer, client or customer

Expert Stakeholders

People and entities that hold AI to account, such as policymakers, legislators, academics and advocacy groups

Product Makers

Companies building or deploying AI-powered products, services, apps and websites

Policymakers

Individuals and organizations developing policy frameworks, principles and requirements at the government level

Considerations for policy

The ongoing democratization and uptake of AI technologies by a wide range of product makers and service providers means more and more people are interacting with AI systems and products every day. As such, policymakers around the world are increasingly advocating for cross-sectoral AI accountability and transparency.

Important policy, regulatory and legislative developments already underway:

The **OECD** "Recommendation on AI" is the first intergovernmental standard for the responsible stewardship of trustworthy AI. Principle 1.3 of the standard pairs transparency and 'responsible disclosure' with explainability: providing 'meaningful information' that is 'appropriate to the context'.

In **Singapore**, the IMDA/PDPC's *Model AI Governance Framework* and *Implementation and Self-Assessment Guide for Organisations (ISAGO)* provide a framework for organizations to operationalize ethical principles for transparency and explainability.

The **European Union**'s *AI Act* takes a risk-based approach towards the development, deployment and use of AI-driven products, services and systems, including user transparency obligations for high-risk systems.

The **United Kingdom** has issued practical advice to organizations to help explain the processes, services and decisions delivered or assisted by AI to the people affected by them.

Despite these developments, the primary focus of AI explainability to date has been on internal engineering and technology development processes. As a result, there is limited understanding of the potential of AI explainability as a vehicle for revealing the intricacies of AI systems to the people who use digital products.

This project embraces this opportunity, exploring practical insights and considerations for people-centric approaches to AI explainability, with general product users in mind.

The insights and considerations contained in this report reflect the explorations that took place in the prototyping workshops, focusing on explainability, transparency, trust and – importantly – control.

As a key theme emerging from this project, control has a complex relationship with explainability.

In the context of AI-powered products, comprehensive control is not necessarily possible or desirable, but some level of control is clearly useful for enhancing people's understanding of AI systems. At the very least, these project findings point to the potential for further explorations of user controls to better appreciate the role of AI explainability for multiple audiences.

While the Framework directly addresses explainability for both general and expert audiences, in a broader sense this project did not focus on the emerging need for greater AI accountability, transparency and explainability for expert audiences. Exploring the explainability needs of expert stakeholders such as academics, legislators, regulators and other third-party organizations tasked with holding AI to account will necessitate a different approach to that taken on this project, which targeted the needs of the majority of people using AI-powered products. Additionally, this project did not consider wider questions about fairness in AI, or explore technical capabilities in detail.

To be effective, AI transparency and explainability policy developments must meet the practical needs of the people using digital products as well as those of the product makers that are required to implement the associated guidance.

Achieving this will require testing, iteration and a cycle of learning both from industry and policy actors.

The investment of TTC Labs and Open Loop in this project reflects Meta's commitment to a continued process of experimentation, sharing and dialog. Through this process we hope to ensure empirical data, practical insights and applied learnings are reflected in emerging product designs and public policy proposals.

How to read this report

This report has **three key sections**, each with a distinct **audience and purpose**

1

AI Explainability Framework

Guidance and design principles for product makers to apply in the creation of AI explainability experiences

2

Product Design Insights

Insights and considerations for product makers to contemplate more broadly in their approach to AI explainability

Product Makers

3

Public Policymaking Insights

Insights and considerations for policymakers to contemplate in ongoing conversations around the development of AI explainability policy guidance

Policymakers

1

AI Explainability Framework

Making AI systems and products more transparent

Meta believes that people who use AI-powered products should have more transparency and control around the collection and use of their data.

This is why transparency and control are core to the company's privacy product outcomes.

In the context of responsible AI, Meta is striving to be more transparent about when and how AI systems support the operation of its products, to make those operations more explainable, and to inform people about the controls they have when engaging with these products. That's why Meta has introduced a number of tools over the years to increase transparency and control.

As part of these efforts, Meta's **Responsible AI team (RAI)** has been developing the **AI Explainability Framework**.

Who benefits from it?

First and foremost, the AI Explainability Framework benefits the people who use AI-powered products. It has been developed with the explicit purpose of helping product makers address the needs and concerns **general product users** have around AI systems and processes.

In doing so, the Framework benefits **product makers**. Effective explainability experiences cultivate understanding and trust, in turn supporting user retention and engagement, and reinforcing the reputation of products as open and transparent.

Expert stakeholders likewise benefit whenever a product maker adopts the Framework, allowing them to better understand these products.

Meta appreciates that general product users and expert audiences are all eager to better understand the ways in which AI systems influence how products work. There are many challenges in explaining the predictions of complex AI systems. Although work in this area is still in its infancy, the hope is that ultimately Meta will be able to build an integrated transparency solution that can automatically feed information from model explainability into new transparency features and controls for people using its products.

Why we need the Framework

People can feel uncomfortable with AI when they don't have an understanding of how, and the degree to which, it affects their experience. The AI Explainability Framework exists to help product makers create AI-powered products with increased transparency, building understanding and long-term trust with the people using their products.

The aim of the Framework is to enable product makers to better understand the various entry points for and factors of explainability. It supports product design decision-making processes by keeping the needs of the people using products at the center of the conversation.

Testing the Framework

The **People-Centric Approaches to AI Explainability** project provided a practical setting to test the AI Explainability Framework with a range of startups. Focusing on the people who use their products, the startups were encouraged to draw on the Framework as a prompt and a reference throughout the Design Jam workshops.

The resulting prototype solutions, featured in the **Product Design Insights**, represent a range of industry use cases in business-to-consumer (B2C) and business-to-business-to-consumer (B2B2C) contexts.

The prototypes validated key aspects of the Framework and helped to identify some valuable opportunities to refine its guidance. Details of these are included in the **Next Steps** section of this report and **Appendix C: Project observations on the AI Explainability Framework**.

How to use the AI Explainability Framework

What is the Framework?

RAI's draft AI Explainability Framework provides direct guidance on the design and development of explainability experiences for AI-powered products.

With a focus on the needs of people using these products, the Framework identifies four dimensions of explainability:

- **AI Awareness**
- **AI Outcome Explainability**
- **AI Product Explainability**
- **AI Model Explainability**

Product design guidelines are provided for each dimension, helping product makers understand people's needs and develop explainability mechanisms to effectively address them.

Designed for ease of use, the Framework's principles and guidance can be put into practice out of the box.

Who is it for?

This Framework is for **product makers** and teams deploying AI-powered solutions.

It has been designed for use by startups and by established companies, in the context of both new and existing products.

It applies equally to digital products and hybrid products, whether AI is fundamental to the service offering or incorporated in a limited or partial way. Product makers can apply the Framework when they are creating explainability mechanisms for new or existing products, or when they are auditing the explainability performance of existing products and services.

In all use cases, the Framework is intended to help product makers explain their AI-powered products to different audiences, with a particular focus on **general product users** and **expert stakeholders**.

How to put it into practice

Product makers can use the **Assessment Questions** to determine which of the four explainability dimensions may be required for their AI-powered products. In most cases, more than one dimension will apply.

These assessment questions are elaborated further in the form of **People Problems**.

Each of the four dimensions incorporates a set of **Design Principles** and **Guiding Questions**, as well as guidance on **Information to Include** in explainability mechanisms.

Product makers can use the **Standard Design Patterns** as templates for commonly occurring explainability experiences.

The Framework also provides **High-Level Design Principles**, an overview of the kinds of **Explainability Experiences** product makers can provide and the specific dimensions they might surface in different **Explainability Touchpoints**.

Framework Dimensions

1 AI Awareness

Information to Include

- Indication of AI involvement in the product
- Confidence level of AI in product outcomes

Audience

- General Product Users

2 AI Outcome Explainability

Information to Include

- Main features / attributions contributing to AI outcomes
- Importance of individual features / attributions
- Presence of human reviewers
- Confidence level of AI in product outcomes

Audience

- General Product Users

3 AI Product Explainability

Information to Include

- Purpose and values of the AI-powered product
- Limitation or risk of the AI system
- Training methods
- Process of how the AI system works
- Input and output of the AI system

Audience

- General Product Users
- Expert Stakeholders

4 AI Model Explainability

Information to Include

- Purpose of the ML model
- Performance of the ML model
- Fairness Evaluation
- Limitation or risk of the ML model
- Input and output of the ML model

Audience

- Expert Stakeholders

AI Awareness
Bring awareness when AI is involved

ASSESSMENT QUESTION

Are people unsure whether their experience is powered by an automated system? Will their confusion lead to negative user experiences?

Information to Include

- Indication of AI involvement in the product
- Confidence level of AI in product outcomes

People Problem

People don't understand how AI is involved in their experiences, which may raise privacy concerns and affect their trust in using the product

Aspiration

Our product informs users if an experience is fully or partially powered by an AI system, building awareness throughout the product experience

Design Principles

1. Surface AI awareness early in the user journey
2. Provide a path to navigate other dimensions of explainability
3. Use intuitive and simple language and visual indications

Guiding Questions

- Do we inform people when they first interact with the AI system?
- Do we use simple language that avoids confusion?
- Do we provide access to additional information and resources?
- Do people have expectations around explainability experiences?

AI Outcome Explainability
Explain individual AI-driven results

ASSESSMENT QUESTION

Do individual AI results present opportunities for confusion?
To what degree will AI errors affect user experiences?

Information to Include

- Main features / attributions contributing to AI outcomes
- Importance of individual features / attributions
- Confidence level of AI in product outcomes
- Presence of human reviewers

People Problem

A negative experience driven by AI feels worse when the user doesn't understand what specifically led to it, especially when it's offensive or unsettling

Aspiration

Our product explains individual AI results and provides people with controls to influence the system

Design Principles

1. Show a set of relevant and clear factors that contribute to the AI result
2. Help people learn more by providing access to product-level explainability
3. Pair with immediate and intuitive feedback mechanisms
4. Communicate the values and impact of user feedback

Guiding Questions

- Do we design explanations in an intuitive and accurate way?
- Are the explanations specific and distinct across different AI results?
- Are the explanations easy to understand regardless of someone's level of technical literacy?
- Do we provide individual outcome-level feedback mechanisms?

AI Product Explainability
Provide an overview of how the AI-powered product works

ASSESSMENT QUESTION

Are people confused about how our systems work? How does this lead to inaccurate assumptions?

Information to Include

- Purpose and values of the AI-powered product
- Training methods (e.g. if privacy-enhancing technologies are incorporated)
- Input and output of the AI system
- Limitation or risk of the AI system
- Process of how the AI system works

People Problem

People don't have a complete or accurate mental model of how the AI-powered product works, which leads to less trust and comfort

Aspiration

Our product provides intuitive ways for people to understand how the AI-powered product works and how their data is used

Design Principles

1. Demonstrate clear product values, capabilities and limitations
2. Break down complex ML processes into intuitive language and visual content
3. Provide stories or examples that demonstrate how the system works
4. Provide product-level controls if possible

Guiding Questions

- Do we clearly communicate the inputs and outputs of the system?
- Do we clearly communicate the benefits and risks of the AI-powered experiences?
- Do we pair product-level explanations with product-level controls?

AI Model Explainability
Provide details about underlying ML models

ASSESSMENT QUESTION

Does the product face regulatory or media pressures?
Are people concerned about the models that power this product?

Information to Include

- Purpose of the ML model
- Performance of the ML model
- Input and output of the ML model
- Fairness evaluation
- Limitation or risk of the ML model

People Problem

External expert stakeholders lack ways to gain a more in-depth understanding of the internal ML models that power the specific product use cases

Aspiration

Our product provides information about the ML models to help people understand model behavior without introducing risks such as adversarial attacks from bad actors

Design Principles

1. Demonstrate how individual model outputs influence the product experience
2. Break down complex ML processes into intuitive language and content
3. Provide easy access to the policies that govern the model's behavior
4. Demonstrate the scenarios in which the model should or should not be used

Guiding Questions

- Do we clearly communicate the inputs and outputs of the model?
- Do we clearly communicate the benefits and risks of the model?
- Do we clearly communicate how the model was trained and evaluated?
- Do we clearly communicate how user data is affected?

High-Level Design Principles

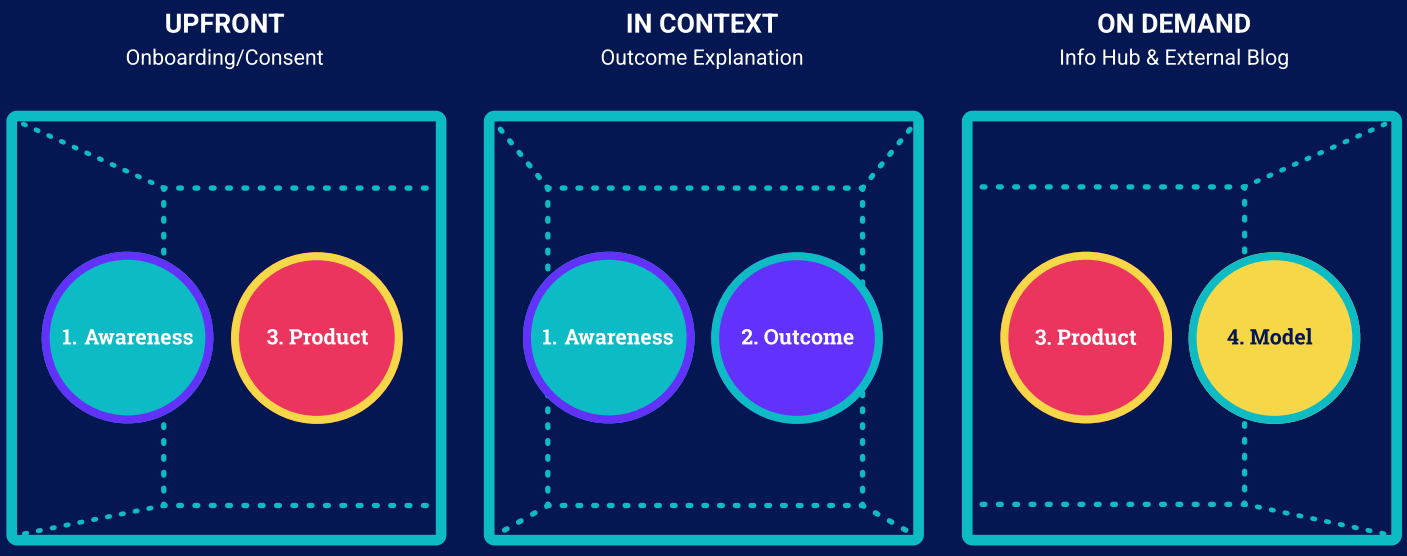
Navigation
Enable navigation among dimensions and ensure they feel connected to each other through consistent visuals and content

Explanations
Ensure users can explore explanations at their own pace and level of depth

Controls
Pair controls with explanations whenever possible, ideally ones related to the same dimension

Explainability Touchpoints

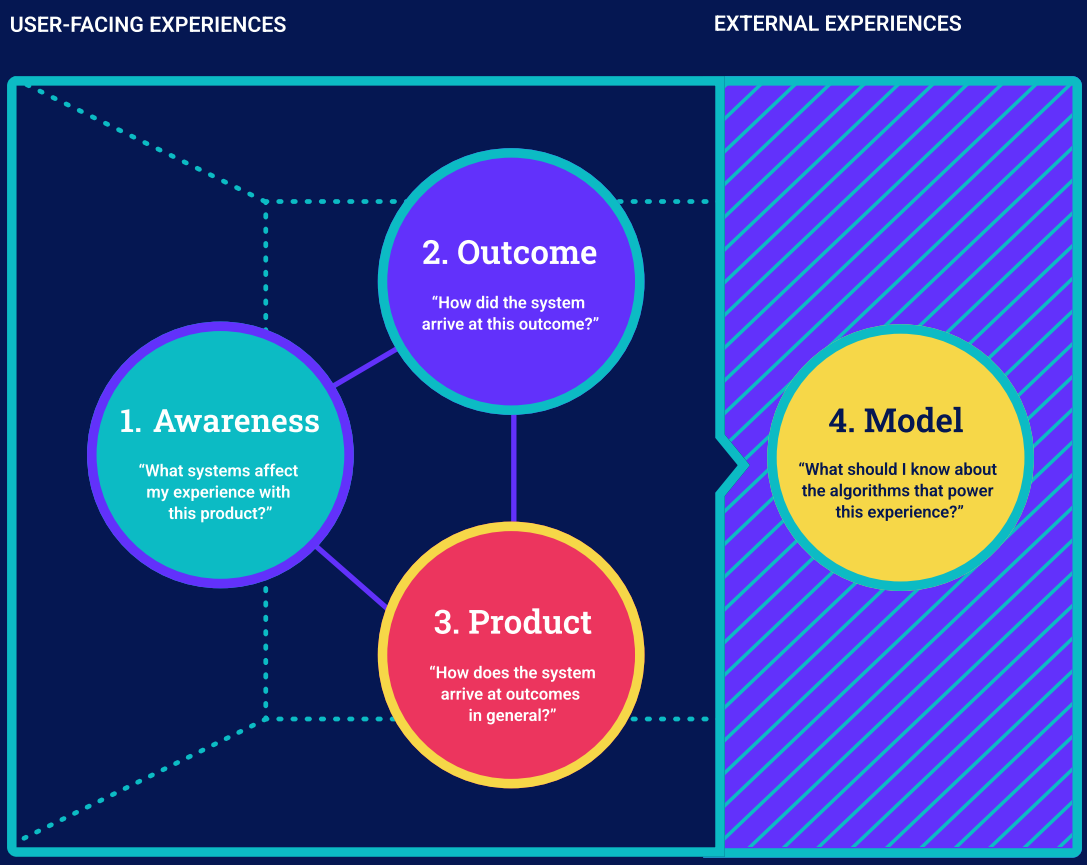
There are three primary ways of providing explainability: upfront, in context and on demand. Representing different stages in a user journey, these touchpoints allow for different kinds of explainability experiences and information, in line with the Framework's four dimensions.



Explainability Experiences

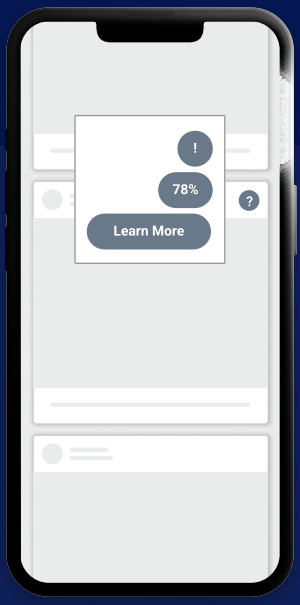
The four dimensions of the Framework are intended to address questions people have in different user-facing and external experiences.

These questions and experiences are interconnected, with information provided at one level often prompting questions at another level. Product makers should provide pathways between different dimensions that allow people to navigate and discover the various levels of information available to them in their user journey.



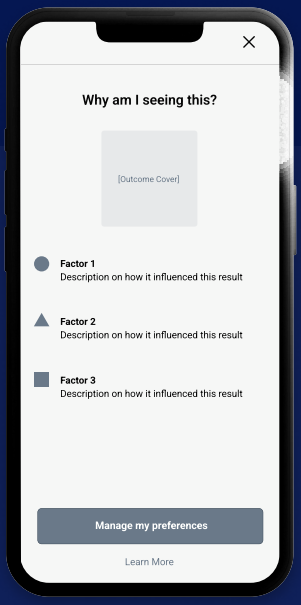
Standard Design Patterns

These templates show a selection of standard approaches for addressing commonly occurring explainability requirements.



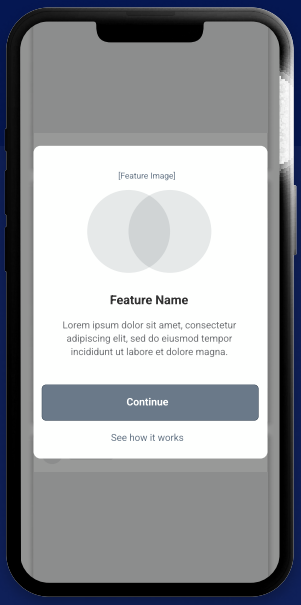
Awareness Sticker

1. AI Awareness



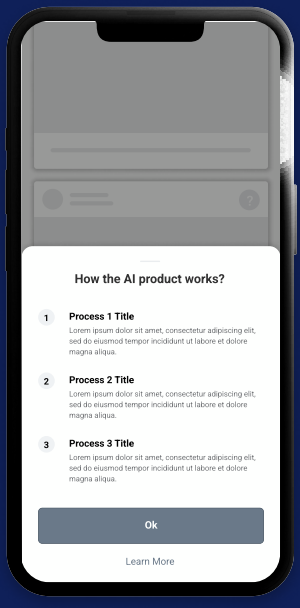
Outcome Explanation

2. AI Outcome Explainability



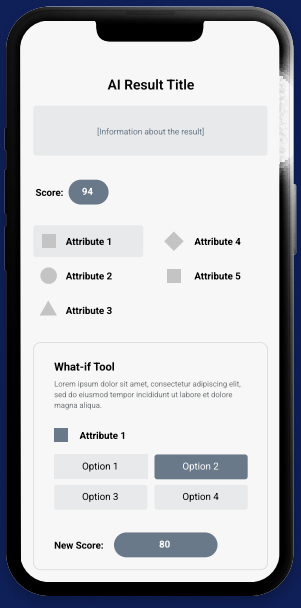
Product Onboarding

1. AI Awareness



Product Explanation

3. AI Product Explainability



What-if Explanation

2. AI Outcome Explainability

2

Product Design Insights

Designing AI explainability experiences to enhance people's understanding

As AI-powered products and features become more common, people are looking to product makers for clarity around these systems and the ways they impact their experience.

When it comes to AI, too much trust can be dangerous. Trust can lead to an over-reliance on algorithms, with people placing their confidence in products that may mislead or cause them harm.

That's why we need more than trust. For the purposes of this current project, TTC Labs has focused on AI explainability that encourages critical thinking, helping people cultivate their understanding of AI systems and their capabilities.

The aim of AI explainability is not to persuade, but to ensure people are informed.

Through understanding, product makers can foster an appropriate level of confidence in their AI-powered solutions.

The main challenge of explainability, then, is finding the most effective ways to build this understanding among diverse users with distinct needs.

There are many audiences for explainability, from product users to external stakeholders and regulatory bodies.

The insights and considerations in this section focus primarily on general users of AI-powered products and services.

Even within this audience, however, we find a diversity of personalities and user types. There are people who want more or less detail, who learn in different ways, who have different levels of digital literacy and trust in AI systems. Some people prefer information to be presented in certain formats, while others want the ability to experiment and take control of their experience.

It's up to product makers to tailor explainability information to meet these different needs.

AI explainability does not discharge a product maker's responsibilities in regards to their AI systems.

In addition to providing AI explanations, product makers are responsible for ensuring these statements are as true and accurate as possible. This includes faithfully representing the capabilities of their systems and any limitations in their predictions.

This is essential for people to place their trust in the explainability information they encounter and to maintain their confidence in AI-powered solutions.

It's also the responsibility of product makers to ensure the information they provide is effective at achieving its aims – that is, helping people understand the implications of using their products and sharing information with them.

While there are technical limitations to the level of clarity that can be achieved around particular AI processes, the focus of people-centric explainability is not the provision of model output data for interpretation and analysis. Rather, it's about making explanations meaningful to people in terms of their experience of a product and how an AI is affecting or influencing that experience.

As part of this, product makers should let users know what safeguards are in place for their protection. Are the AI-driven product results subject to human review? Can people dispute or seek redress for individual results or outcomes they disagree with?

These responsibilities underpin the insights and considerations in this section, providing the basis for improving people's experiences of AI-powered products and features.

How to use this section

These insights have been developed to guide **product makers** in their thinking around the design of AI explainability experiences.

They do not constitute step-by-step instructions, but offer a range of considerations to help identify and prioritize explainability needs, objectives and solutions in particular product contexts.

Both startups and established companies can draw on these insights to create explainability mechanisms for new and existing products and features. They can also be used in the assessment of existing explainability experiences. In all instances, they are intended to help product makers cultivate greater understanding of their AI-powered products, particularly among **general product users**.

The AI explainability prototypes co-created for this project are used throughout this section to explore different aspects of the insights and considerations. They are used together with personas – fictional representations of real users – developed by the multidisciplinary startup teams during the Design Jam, and are supported by further prototype examples of real and fictional apps created for previous TTC Labs Design Jams and Open Loop workshops.

While these various examples focus on in-app experiences, it's important for product makers to also consider external opportunities for explainability, including a product's app store listing, its website and any advertisements.

Product Design Insights

Explainability happens in collaboration

Invite people to engage with explainability information in an active manner through **interactive touchpoints, user controls** and **implicit explainability mechanisms**.

Design is as important as text

Use intuitive design to ensure people can locate, navigate and comprehend explainability information. This includes **common visual language**, **interactive touchpoints** and **standard elements and patterns**.

People need different information at different stages

Support people's developing understanding of an AI by providing the right information at the right time. Identify what they need to know **upfront,** what to tell them **in context** and what they can access **on demand**.

Not everyone requires the same level of information

Determine the depth of information and level of control appropriate to distinct user groups, accounting for differences in the explainability needs of **general product users and expert stakeholders**, of **primary and secondary users** and of **business users and end customers**.

A. Explainability happens in collaboration

People are not passive recipients of AI explanations. They draw on their wider knowledge and experience to develop their understanding of an AI in an active manner.

Different people comprehend AI-powered systems and features in different ways, situating explanations within the context of their experience and updating their ideas based on the way a product behaves and surfaces results.

Understanding this is key to designing people-centric AI explainability experiences. It means adopting a more collaborative concept of explainability, moving beyond the idea that product makers provide explanations and product users merely receive them. It means creating touchpoints that account for different approaches to engaging with information and making AI explainability available in ways that are most meaningful to people.

Product makers can take a more collaborative approach to explainability through the use of **interactive touchpoints**, **user controls** and **implicit explainability mechanisms**.

—

AI explainability isn't one-way. It's an exchange between a person and a product that depends as much on someone's interpretation of an AI – on the assumptions and inferences they make about it – as it does on the information product makers provide.

> "The goal is their understanding, not our explaining.
>
> — Peter Tanham, Meta

Interactive Touchpoints

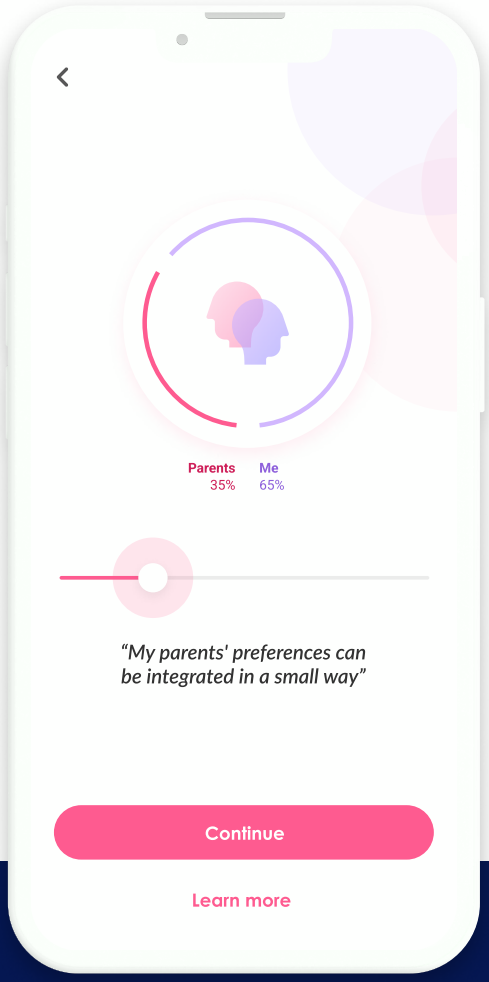
Interactive touchpoints invite people to engage directly with explainability information and learn by doing.

They provide hands-on opportunities for people to discover how an AI is affecting their experience, allowing for the intuitive comprehension of complicated concepts. The more complex the AI, the greater the potential for interaction to cultivate people's understanding.

Betterhalf.ai's prototype demonstrates one of the key possibilities of interaction for explainability. By incorporating explainability directly into the user flow, their solution not only welcomes user engagement, but requires it. Through a simple, interactive slider, the prototype prompts women to determine the influence their parents' preferences have over their suggested matches.

In this example, the slider is central to the setup process, forming a necessary step for women who wish to involve their parents in their matchmaking journey. Offering these users an intuitive way to build their understanding of the AI from the outset, it provides them with valuable context for later interpreting and understanding why they are receiving certain matches.

Not every explanation needs to be interactive, nor can they be. It's up to product makers to determine how to best employ interactive explainability mechanisms based on their unique product contexts and needs of their specific users.



Through a simple interactive slider, women like Priya determine the balance of their own preferences and those of their parents in the matrimonial matches suggested by Betterhalf.ai

As Priya moves the slider left and right, the change in the balance is reflected visually in lines around the icon above, as a percentage split, and in an expressive statement

Priya (35)
Architectural Partner & Director, Mumbai

Persona

"I would love to settle down, but it's important my life partner is the right match"

An independent professional, Priya is responsible for large urban renewal projects locally and across the world. She is facing pressure from her family to settle down and have children, especially as she approaches what her parents consider to be an 'unmarriageable age'.

Priya likes to be able to control her preferences and understand how her data is utilized throughout her digital experiences.

> **Betterhalf.ai is an Indian matrimony matchmaking app that recommends personalized matches with minimal parental intervention**

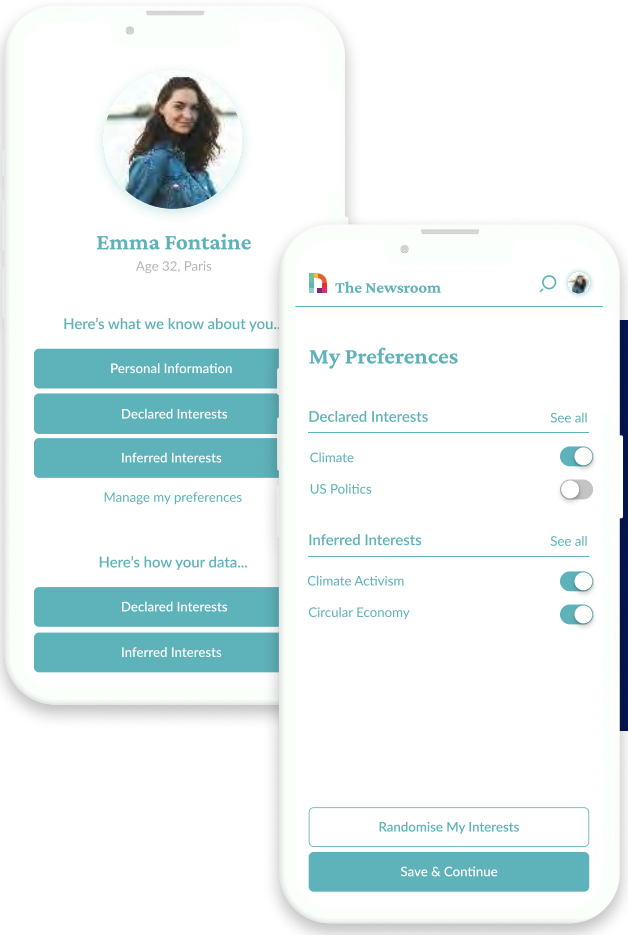
For more detail on the benefits of interactive touchpoints, see the following insight **Design is as important as text**

User Controls

Providing people with limited product controls is an efficient, intuitive way to build understanding and trust.

Controls take many forms. They include mechanisms that allow people to influence the way an AI-powered product generates results, to modify their declared interests and preferences, or to override or remove AI-inferred data inputs. They also include opportunities for people to provide feedback on recommendations and to challenge AI outcomes and explanations if they disagree with them.

Controls can be particularly valuable in the context of collaborative approaches to explainability. When AI is inherent to a product's core offering and there's no way for a person to opt out of AI-powered processes, providing people with some control or influence can reduce the risk of them disengaging entirely.



While providing people with comprehensive control is not necessarily possible, incorporating limited user controls with clear boundaries can greatly enhance people's understanding of the AI and their product experience.

The Newsroom's prototype is a prime example of the effective use of limited controls. Their solution allows people to change subjective inputs into the AI-powered product, including their personal data, declared interests and engagement patterns.

People are not able to influence or override how the AI assesses the trustworthiness of articles, however, as this is fundamental to the product's value proposition. The Newsroom explains how it makes these assessments, but does not provide the option of modifying the process. This distinction enables user control of the newsfeed without compromising the objectivity of the algorithm.

Product makers should consider how changing system variables may impact the accuracy of model outputs and the value people can derive from a service. These trade-offs should be made explicit in the guidance they provide around user controls.

The Newsroom's prototype allows users like Emma to easily manage her preferences

In addition to changing her declared interests, Emma can override her inferred interests: topics the AI has predicted based on her engagement with the app

She also has the option of completely discarding her declared and inferred interests to randomize her newsfeed

> **The Newsroom combats misinformation by curating a personalized newsfeed based on trustworthiness, objectivity and a person's specific interests**

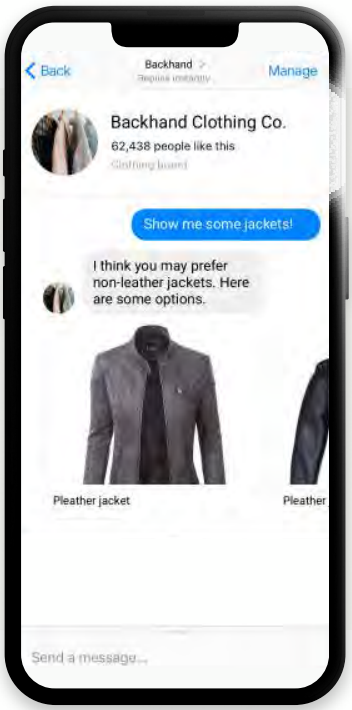
When it comes to controls, product makers need to distinguish between what people can influence and what they can't, and make sure they communicate this to their users.

Similar to The Newsroom, the **Jumper** app provides people with the ability to view and override AI-inferred preference selections.

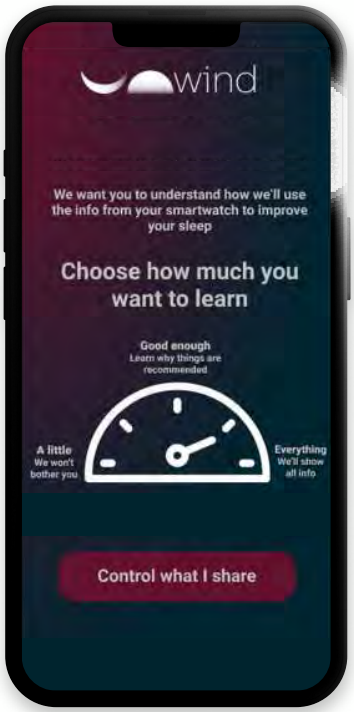
For recommendation engines and matchmaking apps like these, product makers should also consider intuitive ways for people to provide feedback on suggestions. Such feedback mechanisms help people control the recommendations they receive, while also enabling developers to improve their AI systems over time.

Jumper infers a person's interests from their purchase history, suggesting product categories and specific items based on their recent behavior

People can easily access and override these preferences, providing a degree of control over the inputs into the AI's recommendations



> **Jumper is a conversational commerce platform that helps online merchants sell products through messaging and social channels. The Jumper team prototyped this recommendation engine during the first Facebook Accelerator Singapore Design Jam.**



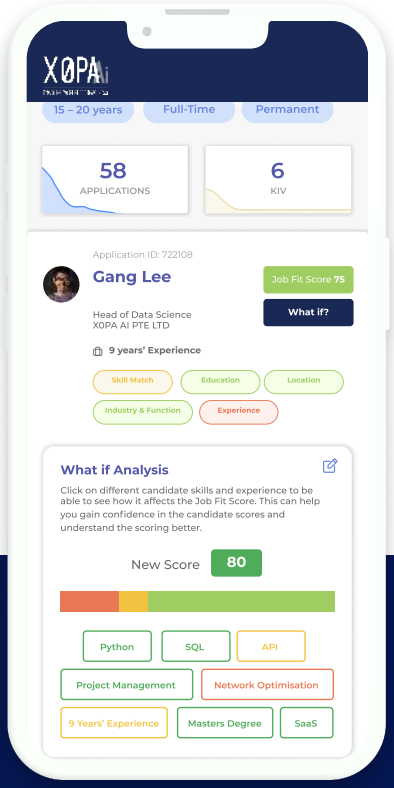
Another app, **SleepRoom**, offers people control over the data they provide and the information they receive.

People choose how much data they want to share from their smart devices to allow SleepRoom to make recommendations

They can also nominate the depth of explainability they want to receive around these recommendations and how their information is being used, from 'A little' to 'Everything'

> **SleepRoom is a fictional digital health service that adjusts bedroom conditions and provides advice to enhance users' sleep. SleepRoom was co-developed in December 2019 at a Design Jam coordinated by Royal Philips, Meta and Considerati Responsible Tech.**

Implicit Explainability Mechanisms



Cathy can use X0PA AI's *What if?* tool to experiment with inputs, changing criteria weightings for skills and experience to recalculate candidate scores and rankings

She is not able to change criteria that could indulge prejudice, such as gender or age, maintaining the objectivity of the algorithm and serving as another example of limited user control

Cathy (38)
Senior Recruiter, Singapore

"I'm wary of using AI to assess candidates for positions"

Cathy is a seasoned recruiter, experienced in traditional recruitment strategies. She's open to new tools if they save her time and optimize her recruitment choices, but she feels she needs a better understanding of the technology before she can confidently engage with these products.

Persona

Explainability information can be considered either explicit or implicit.

Explicit information is typically direct and unambiguous. This type of information can be used to create explainability experiences that are either explicit in form (such as a written explanation), in the way information is provided, or both. Finite and objective, explicit explainability is used to communicate something as clearly as possible, minimizing the risk of misinterpretation. This kind of explainability is essential to cultivating transparency and understanding around AI-powered services.

Implicit information and mechanisms, however, are just as important. A more subtle form of communication, implicit explainability recognizes that people develop their understanding of an AI system beyond the information they encounter directly. It provides opportunities to indirectly build and evolve a person's understanding over time.

Implicit explainability mechanisms allow people to draw inferences based on repeat events and their cumulative experience of a product.

X0PA AI's prototype demonstrates how implicit explainability can work in practice. The *What if?* function creates a counterfactual feedback loop, allowing job recruiters to validate shortlisting and rejection recommendations. By running different scenarios – modifying data inputs and seeing real-time changes – recruiters not only come to appreciate the reasoning behind individual recommendations, but can develop an intuitive understanding of, and confidence in, the product and the AI's role within it.

Typically integrated with the user experience, implicit explainability mechanisms such as this provide the opportunity for people to enhance their comprehension of an AI through regular, everyday use of a product.

Product makers wanting to utilize such mechanisms therefore need to design with them in mind, incorporating implicit explainability into the fabric of their product from the outset.

X0PA AI helps recruiters remove bias and makes hiring more equitable

B. Design is as important as text

AI explainability goes beyond written descriptions, informing people through images, animation, sound, interaction and other non-textual features and elements.

What brings all these elements together is the design of the product interface.

Good user interface (UI) design renders complex explainability concepts clear, simple and accessible for people, regardless of their AI literacy. Product makers should regard interface design as a way to motivate people to seek and discover explainability information, avoiding features that nudge users to skip or ignore important explanations.

Through design and testing, product makers can ensure their interfaces actively cultivate people's understanding of the AI within their products, integrating explainability in a way that meets the needs of different users.

Key features of interface design for explainability include the use of **common visual language**, **interactive touchpoints** and **standard elements and patterns**.

Even with text-heavy explainability touchpoints, presentation is key to ensuring people can access the information they need.

Common Visual Language

Visual language plays a critical role in explainability, showing people how to find explainability information and reinforcing the messages being conveyed.

Familiar visual elements support the usability of explainability touchpoints for both technical and non-technical users. Used in conjunction with standard interaction elements (see page 38), they allow people to draw on their previous experience when it comes to locating, navigating and interpreting information.

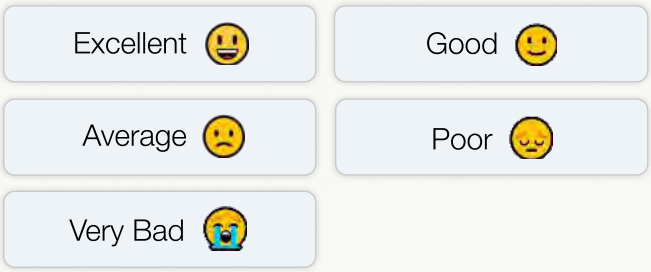
MyAlice's prototype demonstrates how common visual elements can be used to transcend various language and literacy barriers. At the end of a sales interaction, people using the *Product Recommender* are invited to provide feedback, which MyAlice uses to improve the system's AI-augmented purchase suggestions.

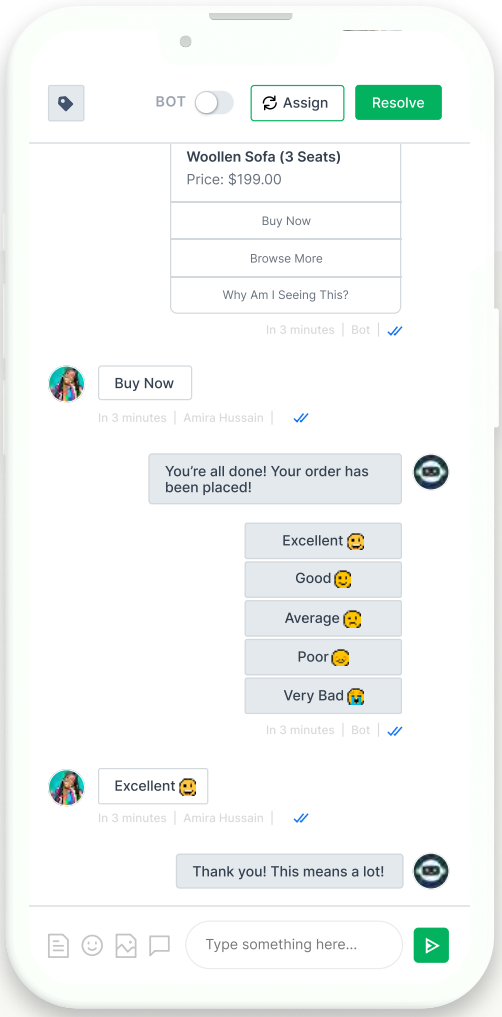
Communicating with people across multiple countries and languages – including customers who code-switch between languages – MyAlice accompanies their feedback options with face emojis. These easily recognizable icons overcome language differences, allowing people to provide feedback quickly and easily.

Product makers need to determine which common visual features to incorporate into their explainability mechanisms.

Customers who use the *Product Recommender* are invited to rate the system

Customers click on one of five feedback options, ranging from 'Very Bad' to 'Excellent', each of which is represented by a different face emoji



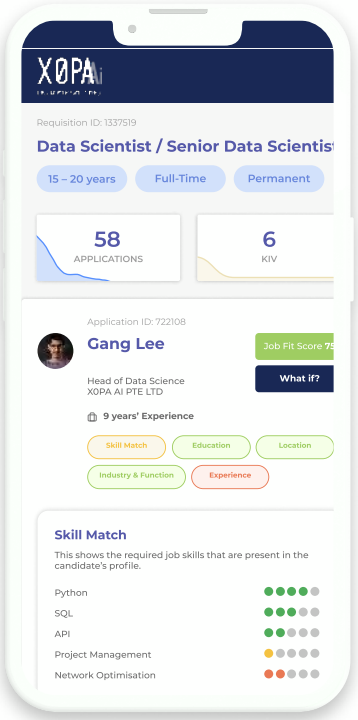


> **MyAlice is a customer support platform for e-commerce owners and sales agents operating across multiple apps and social media channels**

Codes, cues, signs and signals – together with colors, shapes and iconography – are indispensable for cultivating people-centric explainability experiences.

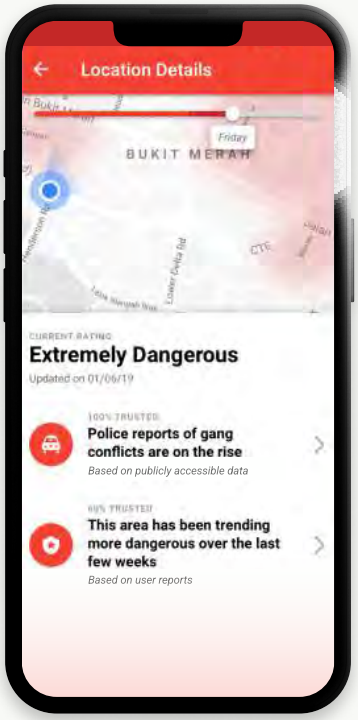
X0PA AI's prototype shows how visual cues can help people understand complex information more quickly. Using traffic-light colors (green, yellow, red), the *What if?* tool allows job recruiters to easily compare candidates and their performance across a range of parameters, as well as understand changes to candidate scores under different scenarios.

Without altering the information provided, this color-coding improves the ability of recruiters to interpret the algorithm's complex outputs and score breakdowns.



A job candidate's performance against different metrics and criteria is color-coded green, yellow and red in the X0PA AI app, allowing recruiters to intuitively interpret scores and rankings.

> **X0PA AI helps recruiters remove bias and makes hiring more equitable**



Common visual language supports intuitive wayfinding for both new and existing users.

Another example of common visual language, the fictional app **Sayfe**, shows the value of familiar features when the underlying concepts are difficult to grasp. Borrowing from weather forecasting services, Sayfe adopts familiar visual concepts such as heat maps to visualize AI predictions.

This intuitive use of visual language is particularly useful in the context of a safety app, allowing users to quickly comprehend dynamic and evolving information.

Sayfe uses a heat map to indicate the likely level of safety in a particular area, with progressively darker reds showing increasing levels of danger

People can turn data sources on and off using layer controls similar to those in standard map and navigation apps

> **Sayfe is a fictional mapping and transit app designed to help people avoid areas that it predicts are or may become a public safety risk**

Interactive Touchpoints

Interactive touchpoints communicate complex concepts more efficiently than written explanations.

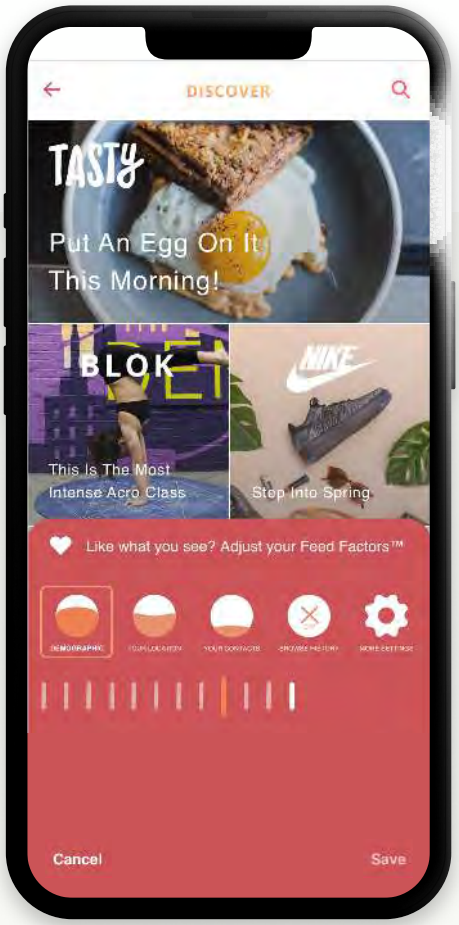
Just as interactivity is a key consideration for product makers adopting a more collaborative approach to explainability, it plays an important role in the design of explainability mechanisms. Intuitive design supports the use of interactive elements, encouraging people to learn through doing.

Interactive interfaces generally avoid overwhelming people with comprehensive technical information. Instead, they invite people to develop their understanding iteratively and incrementally based on the outcomes of their actions.

The fictional app **Loco**, for example, includes a set of interactive dials for controlling the algorithm's inputs. Playing with these dials allows people to understand how their personal data influences the content that Loco is surfacing in their feed.

Interactivity not only supports certain kinds of learning modalities – specifically, kinesthetic learning styles – but it also meets the evolving expectations of product users. People want to actively participate in their learning, not just read generic statements. Interactive elements allow users to direct the pace and timing of their exploration and understanding.

Product makers need to decide how interaction can be most effectively incorporated into the design of explainability touchpoints and user flows.



The Loco prototype features four adjustable dials that control a user's Feed Factors: Demographic, Location, Contacts and Browser History

People can use these dials to amplify or diminish the influence of the different data points on their feed – or turn them off completely

> **Loco is a fictional social app that allows people to share images, videos and augmented reality (AR) masks directly with friends or to a feed**

For more detail on the benefits of interactive touchpoints, see the previous insight **Explainability happens in collaboration**

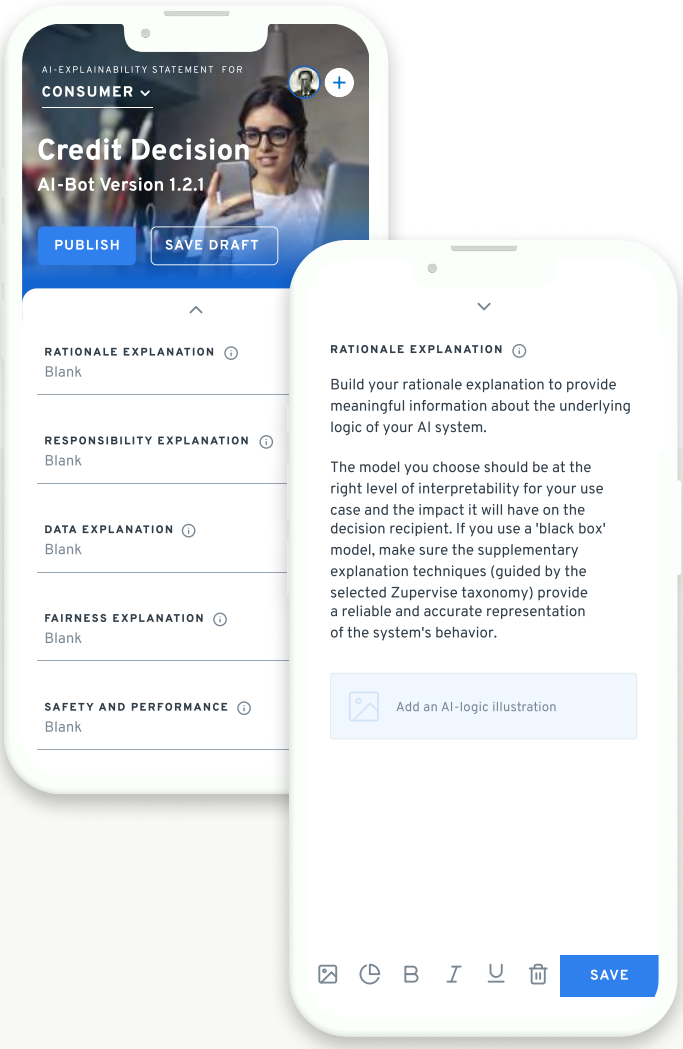
Standard Elements & Patterns

The reason common design patterns recur across products and services is simple: people know how to engage with them.

Product makers can incorporate standard elements, formats and layouts to provide consistent explainability experiences within a product or across a suite of products. Considerations around common design patterns include how information is displayed on an interface, as well as the elements people use to navigate or trigger actions within a product, such as buttons, tooltips and input fields.

Zupervise's prototype draws on a range of standard elements to demystify the process of creating AI explanations. Through an established interaction vocabulary – dropdown menus, expand/collapse headings, tooltips – product makers can effectively tailor explainability statements for different audiences. At the same time, familiar authoring, editing and review features enable technical and non-technical contributors to collaborate in a meaningful way.

Familiar design features greatly enhance a person's ability to process complex concepts and integrate them into their understanding.



Having selected the target audience (e.g. consumer) from a drop-down menu in the Zupervise interface, Ricardo can begin populating the explainability statement's required fields

Clicking on a field such as Rationale Explanation allows Ricardo to write, edit and format the information relevant to that field, adding images or video content as required

Ricardo (44)
Risk Lead (Financial Services), London

" It's my job to ensure the organization is protected against AI risks"

Persona

Ricardo leads the risk office at a challenger bank. When the business leverages AI and machine learning algorithms, he needs to satisfy internal risk management requirements as well as the explainability requirements of diverse stakeholders.

Zupervise is a unified risk transparency platform to govern AI in the regulated enterprise

C. People need different information at different stages

Getting the right information at the right time is fundamental for people to understand how an AI is impacting their product experience.

Throughout their experience of a product, people require different pieces of information to develop a coherent, holistic picture of an AI system. The distinct explainability experiences that people have in different moments are key to building their overall comprehension and understanding.

When it comes to designing explainability and structuring these experiences, product makers need to identify the specific stages and junctures where information will be surfaced, together with the types of information that will be provided at these touchpoints.

A useful way of breaking this down is to determine what to make available **upfront** and **in context**, and what people will have access to **on demand**.

This distinction between upfront, in-context and on-demand information draws on previous **TTC Labs research around data disclosures** and aligns with the explainability touchpoints in the **AI Explainability Framework** included in this report.

Upfront Notifications

A product's sign-up and onboarding process provides the opportunity to introduce core explainability concepts. This is where people typically find out if a product is using AI and the reasons why, learning how algorithms are both enhancing and affecting their experience.

Product makers can incorporate upfront notifications (such as notice and consent mechanisms) into hold points and moments of friction at this stage of the user journey, prompting people to engage with explainability information.

While upfront notifications are often provided when people first open an app or start using a feature, they can also be made available prior to this, in product descriptions or promotional information.

Betterhalf.ai's prototype demonstrates how upfront touchpoints can be used to build AI understanding from the outset. During the setup of the *Parental Preference* feature, women are invited to nominate the comparative weightings of their preferences and those of their parents.

Learning how to set these preferences informs women about the impact different inputs have on the matchmaking process. Betterhalf.ai thus encourages user engagement with explainability information by integrating it with general product guidance. In doing so they help cultivate their users' understanding of the AI and their trust in the recommendations it makes.

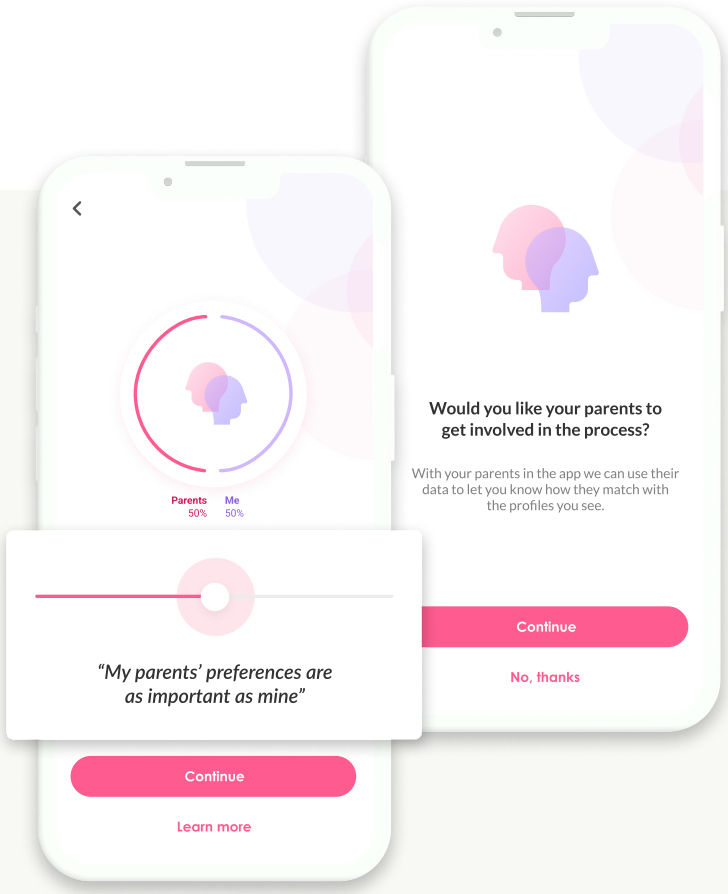
Upfront touchpoints are where explainability conversations start, not where they end.

Product makers need to consider creative ways of giving people enough information to make informed decisions, without inundating them with detail.

Through a consent screen, women decide whether they want to involve their parents in the matchmaking process

If they choose to continue, users are prompted to set the balance of their own preferences and those of their parents through an interactive slider

The slider screen features a 'Learn more' button that provides users with a more in-depth explanation of how the slider influences matches before they begin the matching process



> **Betterhalf.ai is an Indian matrimony matchmaking app that recommends personalized matches with minimal parental intervention**

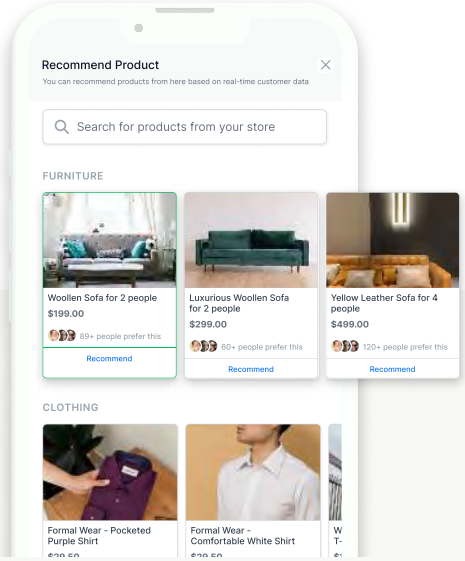
In-Context Explainability

Contextual explainability mechanisms inform people about AI outcomes and results as they are generated.

They offer insights into specific AI processes and algorithmic outputs in the moment, allowing people to interpret explanations in the context of their immediate experience.

Whereas upfront notifications tend to be one-off moments, in-context mechanisms have the benefit of providing regular, and often repeatable, experiences. They allow people to develop and deepen their understanding of an AI and its processes over time.

In-context touchpoints also allow product makers to reach all their users, not just new ones, ensuring explainability information is available and accessible to anyone using their product.



MyAlice is a customer support platform for e-commerce owners and sales agents operating across multiple apps and social media channels

The key with contextual explainability is knowing when it should interrupt a user flow and when it should be in the background, ensuring these mechanisms are visible and available while supporting the overall user experience.

A good demonstration of in-context explainability is provided by **MyAlice**'s *Product Recommender*. This prototype solution supports the ability of sales agents to decide which recommendations to make to an end customer by explaining why the AI is suggesting particular products. Relevant evidence is then passed on to the customer, helping them understand how the AI and the agent made a particular recommendation.

This prototype demonstrates how the same explainability mechanism can serve the needs of different types of users in the same interaction. Focused on the specific needs and interests of the customer and their evolving conversation with the agent, this is an example of explainability information that can only be provided in context.

The system surfaces recommendations in response to a customer inquiry (e.g. sofas), drawing on data such as purchase history and sales of similar products

Sales agents like Roy see the social proof behind the AI's selections, including the number of people who bought a particular product and the average rating they gave it

Customers can then access this social proof through a 'Why Am I Seeing This?' button, using the information to understand how the AI (and Roy) made the particular recommendation

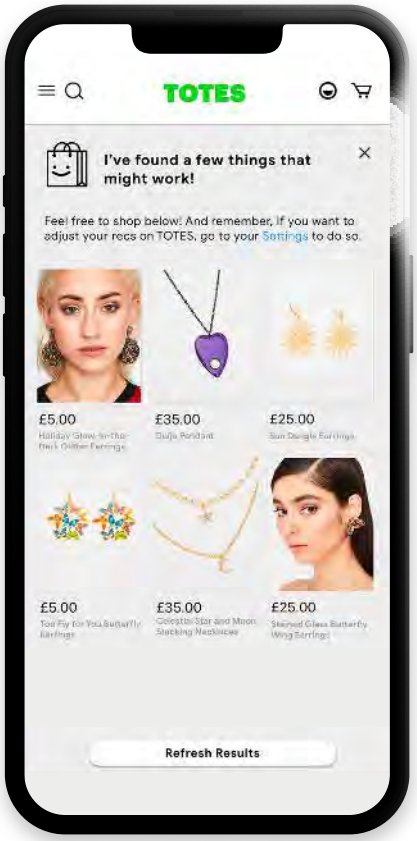
Roy (24)
Sales Agent, Vietnam

"I'm interested in anything that can help me work faster and more effectively"

Persona

Roy is incentivized to respond to customers quickly and accurately, so speed is key. As is customer satisfaction. Roy receives bonuses if he can resolve customer queries positively and convert them to sales in less than 15 minutes.

For existing products and apps, product makers should think beyond new users, introducing explainability information to all users through engaging product experiences.

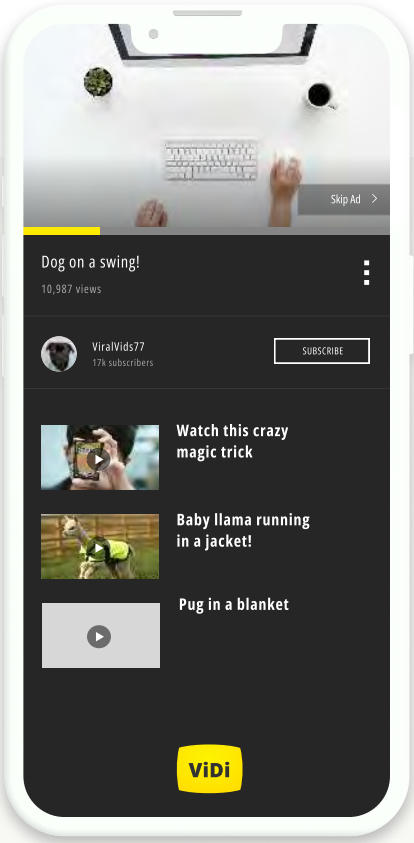


Totes makes explainability available to new and existing users through in-context notifications.

When people receive personalized product suggestions they also see a notification letting them know how to adjust their preferences

Providing these notifications in context reminds all users that they can actively curate their product experience

Totes is a digital marketplace that sells a wide range of goods from thousands of brands worldwide



An alternative example, the fictional app **Vidi**, allows people to change recommendation inputs in context, without navigating away to a settings screen.

Weekly preference notifications keep parents up-to-date with any AI-inferred interests added to their child's profile

Parents can remove any unwanted topics from their child's profile on this screen, with changes taking effect in real-time.

Vidi is a fictional video-sharing platform that creates a feed based on the viewing behaviors of teens and their friends

On-Demand Information

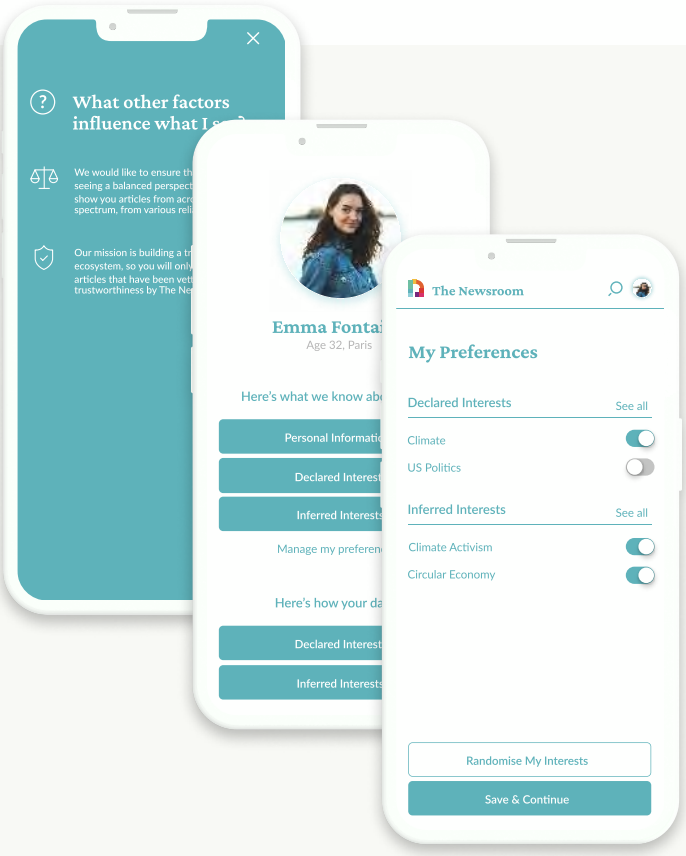
More isn't always better. Providing as much detail as possible – even in an on-demand explanation – doesn't necessarily create greater transparency.

On-demand explainability is generally available for people to access at their convenience, rather than during a specific action or at a designated stage. Like contextual mechanisms, on-demand information is intended for both new and existing users.

On-demand explanations need to cater to different levels of AI literacy and different explainability appetites, balancing the desire of some people for in-depth information with the needs of others for simple, concise descriptions. The aim is not to explain everything, but to maximize people's understanding so that they can take action, where necessary, and make informed decisions.

The Newsroom's prototype provides transparency around the curation of people's newsfeeds by explaining both the objective and subjective inputs into the personalization algorithm. The objective inputs include trustworthiness assessments and political balance, while the subjective inputs include a person's declared interests and preferences, as well as the interests the AI has inferred based on their engagement patterns and past behavior.

Product makers should determine the best place for on-demand information – whether this is within a product or in an external location, such as a product website or an after-service communication.



The Newsroom combats misinformation by curating a personalized newsfeed based on trustworthiness, objectivity and a person's specific interests

- Users like Emma can access explanations on the factors influencing the product in a general sense
- Clicking through to her profile, Emma can view the inputs the algorithm uses to personalize her newsfeed: personal information, declared interests and inferred interests
- Emma can then manage her preferences around these inputs, toggling topics of interest on and off or randomizing her feed

Emma (32)

Technology Worker, Lisbon/Paris

"I'm extremely sensitive about platforms using my data"

Persona

An avid news reader, Emma uses The Newsroom to stay informed and unbiased, allowing her to confidently engage in conversations with people who think differently from her. She's aware of data and privacy issues, wanting to know how her data is being used.

D. Not everyone requires the same level of information

Highly detailed explanations aren't always useful, desired, or even possible. The key to creating positive user experiences is knowing how much to reveal and when.

People who require explainability about AI-powered products divide into different audiences. These audiences are defined by the kind of engagement they have with these products and services, and their specific explainability needs regarding them.

At a high level, explainability audiences can be divided into **general product users and expert stakeholders**. For the purposes of this report, a general user is considered to be anyone who uses an AI-powered product or service, either in a professional or personal sense. Expert stakeholders, by contrast, include policymakers and regulatory bodies, media, labor organizations and advocacy groups.

Within the category of general product users, people can be further classified along the lines of **primary and secondary users** and **business users and end customers**.

These audiences are a crucial consideration in the design of explainability experiences. Product makers need to ensure people are provided with the information most relevant to their respective needs and contexts.

They also need to consider how explainability is balanced between different audiences, acknowledging the trade-offs that come with this. When a product serves different user groups, the information and control appropriate to each can vary significantly.

—

These audience divisions are by no means comprehensive. Product makers need to identify the most appropriate ways to classify their specific audiences and target explainability information accordingly. These observations are explored further in the section **Effective policy takes product makers into consideration**.

> **"Take planes, for example – you don't need to explain everything for someone to trust them**
>
> — Sang Hao Chung, PDPC

General Product Users & Expert Stakeholders

Explainability is not one-size-fits-all. General product users have different needs to expert stakeholders such as practitioners and regulators.

The respective needs of these audiences are not fixed, but change in relation to different product contexts. Both general users and expert stakeholders have distinct requirements based on the service being provided, the role of AI within the product, and the industries and jurisdictions where the product maker operates.

Provide people with information specific to their context, rather than trying to cover off everyone's needs with the same explanations.

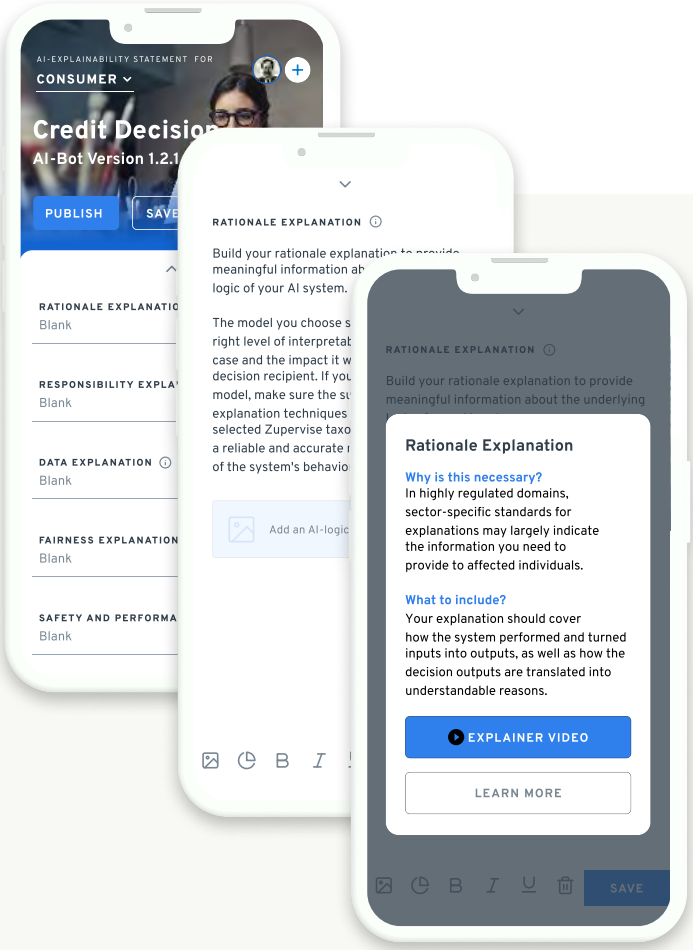
Zupervise's prototype addresses this challenge by enabling product makers to produce explainability statements for different audiences. Designed for teams deploying AI-powered products and services, the platform offers a range of tools and templates for tailoring explainability information to the diverse needs of general product users (consumers) and expert stakeholders (regulators and practitioners) as required. Each audience receives an explainability statement specific to their context, without being overloaded with information intended for others.

Product makers need to understand who their explainability audiences are and get to know their specific issues around trust, understanding and confidence in AI. This will allow them to make the most meaningful information available to each group and best satisfy their diverse needs.

Product teams using Zupervise select a target audience from a drop-down menu, then build out a statement with report templates and categories matched to the needs of this audience

Common elements and fields such as 'Rationale Explanation', 'Responsibility Explanation' and 'Fairness Explanation' allow them to take a structured approach to explainability

The platform provides audience-specific guidance for creating explainability information, including written explanations and short explainer videos



Zupervise is a unified risk transparency platform to govern AI in the regulated enterprise

Primary & Secondary Users

What happens when the explainability needs of one group of general product users conflict with those of another? This is not just about different levels of AI literacy. It's about balancing competing interests and understanding the trade-offs involved, especially when meeting the needs of one kind of user makes it impossible to meet the needs of another.

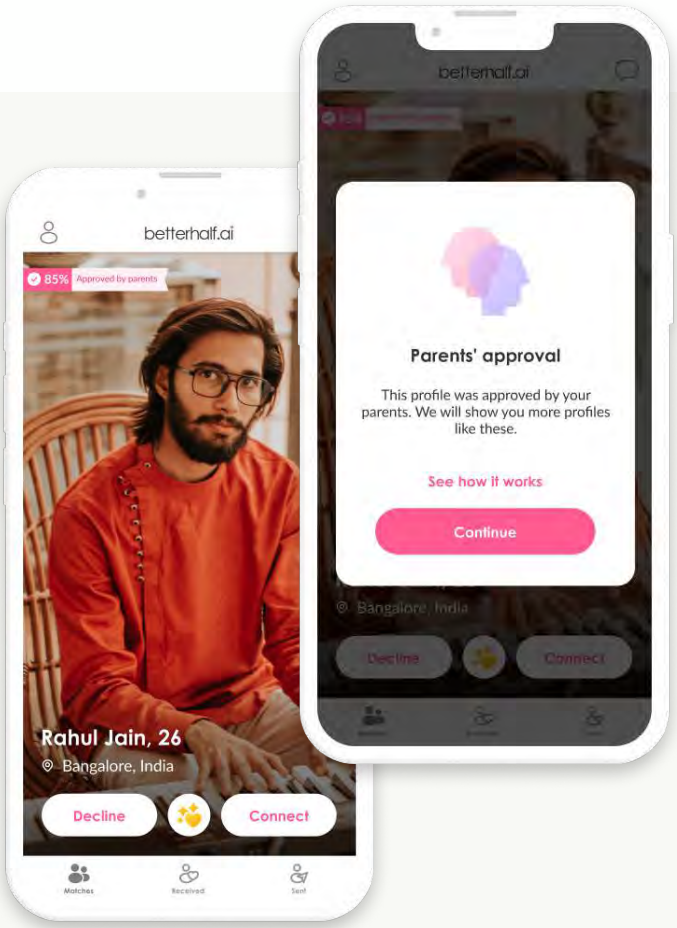
Categorizing users into primary and secondary groups provides clarity around whose interests take precedence in different contexts.

Betterhalf.ai's prototype demonstrates an innovative approach to this problem, positioning women as primary users and their parents as secondary users. If a woman chooses to include her parents in her matchmaking journey, she is prompted to nominate the influence their preferences have over her recommended matches.

On a practical level, this approach means the control needs and explainability requirements of the parents are subordinate to those of the women using the app. As secondary users, parents have limited control and visibility over the AI recommendations. Instead, women retain control over their matchmaking and dating lives, with the ability to modify the impact of their parents' preferences on the AI, the ability to ignore their selected matches and, ultimately, the ability to choose which matches they pursue.

Priya, as a primary user, nominates the balance between her preferences and those of her parents

She can then see the influence her parents' preferences have over individual matches she receives – information to which her parents, as secondary users, do not have access



Betterhalf.ai is an Indian matrimony matchmaking app that recommends personalized matches with minimal parental intervention

Persona

Priya (35)
Architectural Partner & Director, Mumbai

"I would love to settle down, but it's important my life partner is the right match"

An independent professional, Priya is responsible for large urban renewal projects locally and across the world. She is facing pressure from her family to settle down and have children, especially as she approaches what her parents consider to be an 'unmarriageable age'.

Priya likes to be able to control her preferences and understand how her data is utilized throughout her digital experiences.

Business Users & End Customers

When products service businesses as well as their end customers, different explainability may be appropriate to these distinct types of general product user.

Business users typically want some degree of control over AI-powered products, as they are seen to share at least some responsibility for the system. As far as their end customers are concerned, business users are accountable for the outcomes and results surfaced by the product. But providing this control to business users does not discharge the product maker's responsibilities. Rather, it requires clear communication with these users around the product's capabilities – what it can and can't do, and how much confidence they can place in the system's predictions and performance.

X0PA AI demonstrates some key differences in the explainability information and control that might be made available to business users (job recruiters) and end customers (job applicants). While both groups are provided with information on how the AI system generates results, only the recruiters can access details regarding specific computations made by the AI-powered product.

As with non-automated recruitment processes, it's at the discretion of the recruiter to provide individual shortlisting and rejection information to applicants.

This approach highlights a situation where transparency for one group (business users) is greater than that provided to another group (end customers) – a trade-off in the effort to balance their competing interests.

For the same reason, candidates do not have access to the counterfactual *What if?* tool. Designed to help recruiters validate algorithmic outcomes, providing this functionality to applicants would potentially expose recruiters to challenges and disputes.

Product makers need to satisfy different user needs without compromising their IP, their commercial obligations or their core product offering.

The Job Fit Score and score breakdown shows Cathy how X0PA AI has assessed a candidate's suitability for a role

Cathy can then use the *What if?* tool to explore which criteria might improve a candidate's scoring under different scenarios (e.g. removing location from the assessment)

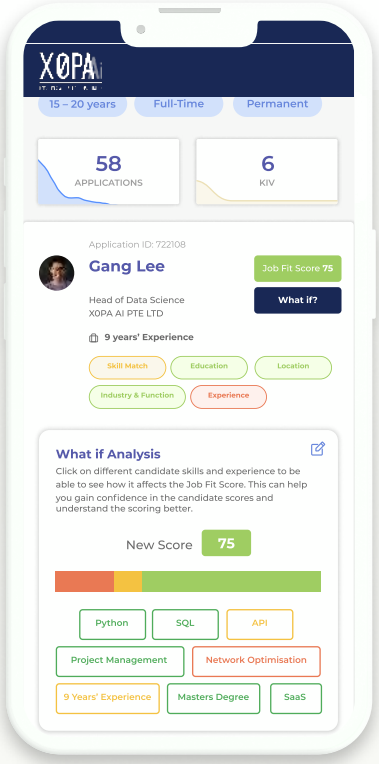
Job candidates do not have access to their Job Fit Scores, score breakdowns or the *What if?* feature

Cathy (38)
Senior Recruiter, Singapore

Persona

"I'm wary of using AI to assess candidates for positions"

Cathy is a seasoned recruiter, experienced in traditional recruitment strategies. She's open to new tools if they save her time and optimize her recruitment choices, but she feels she needs a better understanding of the technology before she can confidently engage with these products.



X0PA AI helps recruiters remove bias and makes hiring more equitable

3

Public Policymaking Insights

Creating AI explainability policy guidance to improve people's experiences

As people look to better understand the inner workings of AI-powered products, there is a growing need for policymakers to design effective, actionable AI explainability policy.

Within the broad field of AI policy and governance, transparency – and in particular, explainability – is a relatively new area of expertise.

Policymaking approaches to AI explainability, whether through normative requirements, operational guidance or technical standards, need to account for technology, use cases, contexts, applications, ideas and inputs that are continuously and rapidly evolving.

Policymakers can address this complexity by factoring in risk and impact to determine when explainability is required and to what extent. In this way, policy can reflect the contextual nature of explainability. Rather than taking a one-size-fits-all approach, such risk-based requirements take into account the actual need for explanations based on specific types of AI applications, their intended purpose, and the impact they will have on the people using or affected by them.

To do this, policymakers need to know what actions, on the part of product makers, will lead to people gaining a better understanding of AI systems.

The question, then, is how policymakers can uphold the interests of society, promoting responsible and ethical approaches to AI through sound explainability policy guidance, without creating impractical or onerous requirements that stifle innovation.

What does AI explainability policy guidance need to consider, what shape might it take and how should it be created?

How to use this section

These insights have been created to support policymakers in the development and implementation of AI explainability policy grounded in product and technical considerations.

They were developed with a focus on product makers and their needs (*who*), the various forms a policy can take (*what*) and the ways in which policy is created (*how*).

Policymakers involved in the development of frameworks, principles, standards or requirements at a government level can draw on these insights and considerations to support their work. They are intended to help these policymakers identify opportunities to work with product makers to create and operationalize policy that delivers people-centric AI explainability experiences.

Policy can be instantiated in different formats and pursued through different instruments, from laws and regulations to principles, standards and governance frameworks. Rather than looking at the specific features of these individual formats and instruments, however, this section focuses on policy at the higher level of normative guidance. Here, the term *policy* is used in a versatile manner, signifying a set of ideas that guide and shape behaviors – policy as a basis for making decisions and taking action.

The focus on product makers in this section results from the particular scope of this project and the central role product makers play in the creation of AI explainability experiences for general product users. This is not to dismiss or undermine the roles of other policy users and stakeholders, such as advocacy organizations, civil society representatives and academics. These stakeholders are vital in supporting and shaping the role of policymakers in defining, developing and implementing AI policies.

A. Effective policy takes product makers into consideration

Craft policy that addresses the needs, contexts and challenges of product makers implementing AI explainability by **thinking of product makers as policy users**, **identifying different AI explainability audiences** and **understanding the explainability needs of these audiences**.

B. Adapting form and content provides entry points into policy guidance

Modify policy documents to make them more relatable and actionable for product makers, **supporting the alignment of their ethical commitments with policy guidance**, **helping product makers find what they need within policy guidance** and **making policy clearer and more explicit**.

C. Collaboration drives better policy outcomes

Bring together policy and product teams to realize the shared ambitions for people-centric explainability, through **closer alignment of policy and product governance**, **supporting the product development process** and **reimagining the relationship between policy and product**.

A. Effective policy takes product makers into consideration

This means understanding the challenges for product makers in implementing AI explainability and crafting policy with their needs in mind.

Traditional policy, as a generally applicable set of normative provisions or requirements, tends to have a horizontal outlook, focusing on a set of applications, an array of technologies, a group of businesses or an industry as a whole.

This project has highlighted, however, that in addition to reaching across the ecosystem of AI-powered and data-driven services, explainability policy also reaches vertically, to specific roles within companies. These are the individuals responsible for interpreting and applying policies within a business – the people whose particular needs and contexts impact how faithfully a policy can be put into practice.

The observations in this section align with key product design considerations, as detailed in the Product Design Insight **Not everyone requires the same level of information.**

Thinking of product makers as policy users

What if

... policy were written for a range of policy user personas? Are there common roles for which these could be created in terms of AI explainability and transparency?

Just as products have users, so do policies. And just like product users, policy users come in different shapes and sizes.

Considering product makers as policy users can help identify their distinct policy needs and AI explainability requirements, including any challenges they face when implementing policy.

One way to approach this is to borrow a tool from design thinking: the persona. A persona is a fictional character that represents an end user or stakeholder. Personas are used to foster empathy for a person's attitudes, beliefs and concerns.

The policy user persona shown here has been adapted from the product user persona developed by the **Zupervise** team during the Design Jam. It shows how personas can help policymakers articulate and understand the specific needs and challenges of different product makers, enabling them tailor guidance to these particular policy users and their contexts.

A note on policy users

While there are a number of important policy users and stakeholders, these insights and considerations focus primarily on product makers as policy users.

Ricardo (44)

Risk Lead (Financial Services), London

"I need to be able to communicate requirements, best practice and expectations to the bank's internal stakeholders"

Policy User Persona

Accountable to the executive and the board of a challenger bank, Ricardo seeks to institute explicit processes for validating and approving the design, development and deployment of ML algorithms.

Ricardo's needs and challenges with AI explainability policy are:

- Maintaining clarity around policy requirements despite ongoing regulatory changes
- Communicating AI explainability requirements to internal teams and leadership
- Understanding how AI technology is being developed, updated and deployed internally

Identifying different AI explainability audiences

What if

... policy provided guidance on identifying product makers' different explainability audiences and stakeholders?

Product makers need to provide transparency and explainability to a range of audiences.

These audiences are determined by the type of services the product maker offers, the role of AI in their products and the industries and jurisdictions they operate in.

While some of these audiences may never directly engage with the policy guidance at stake, they remain the intended beneficiaries of its provisions.

Realizing effective policy outcomes for these audiences requires product makers to implement guidance with regard to diverse explainability use cases. To do this, they first need to know who their audiences and stakeholders are.

- General Product Users
- Unions & Labor Organizations
- Product Maker Teams & Corporate Divisions
- Civil Society & Advocacy Groups
- Senior, Executive & Board Representatives
- Media
- Parents & Guardians
- Government & Regulatory Bodies

> *An upcoming Open Loop report will explore audiences in more detail, connecting them to the different contexts, purposes and content of AI explainability solutions.*

Understanding the needs of different audiences

What if

... policy helped product makers adapt and tailor explainability information to different audiences?

Because people's needs are contextual, so is explainability.

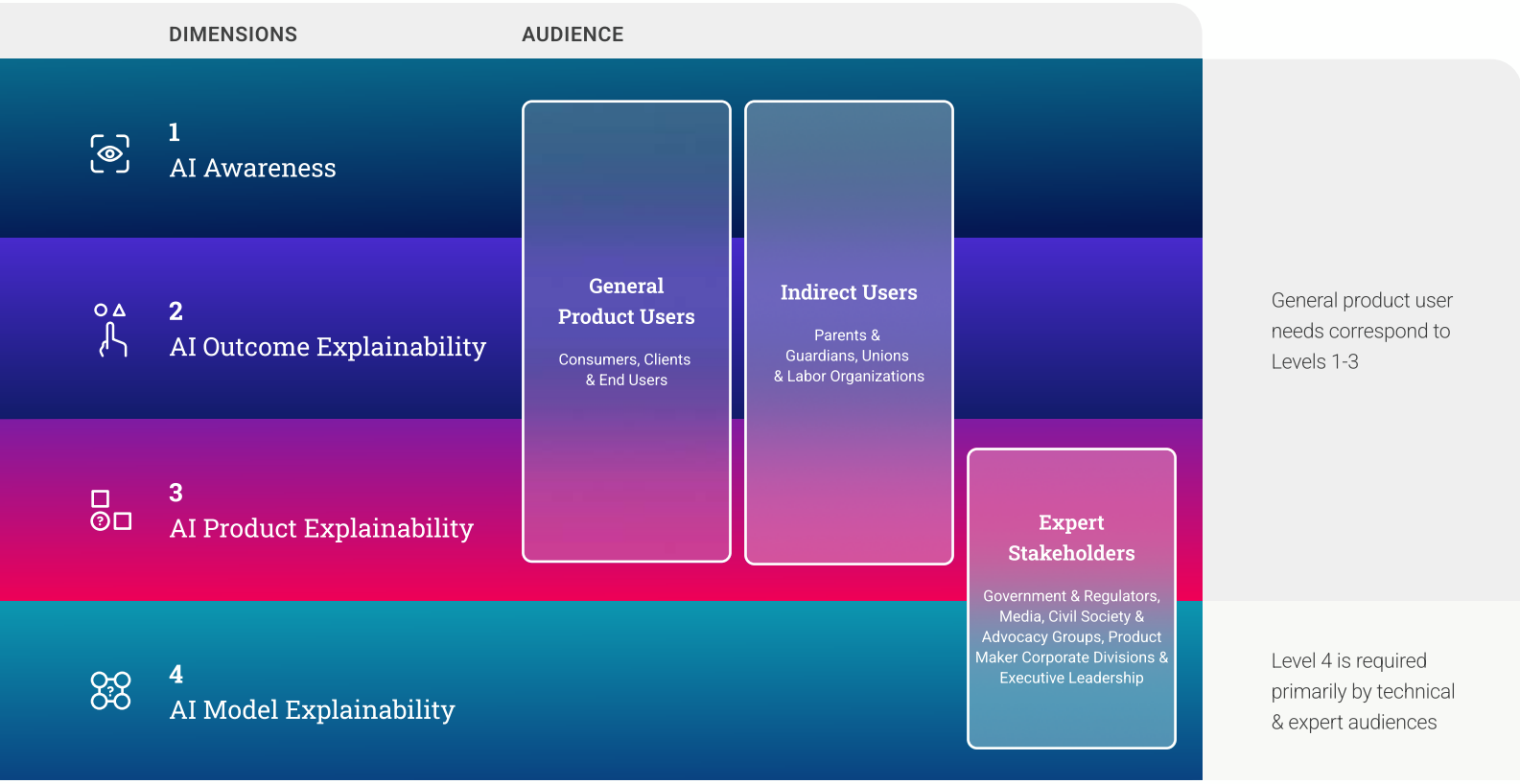
Different use cases give rise to distinct requirements and expectations for different audiences. What might be considered sufficient information in one context could be either excessive or vastly inadequate in another. For example, a jobseeker using **X0PA AI**'s recruitment service requires different assurances to a woman looking for a partner through **Betterhalf.ai**'s matrimony matchmaking app.

But while the explainability information required by these audiences differs in content and by degree, it is still of the same kind.

Testing the draft **AI Explainability Framework** developed by Meta's Responsible AI team (RAI) has revealed that different audiences, despite their diversity, often require the same kinds of explainability information – whether that's being made aware there is AI in a product, understanding how that product generated a particular result, appreciating how the product works in more detail or assessing the performance of the AI model.

What sets them apart is the depth of information they require and the way product makers need to formulate explanations to address their specific contexts.

A policy may never be able to advise on the exact content of individual explanations, but it can provide broad guidance, framing entry points and providing high-level orientation based on audience types. Policymakers have an opportunity to help product makers understand the needs of their different explainability audiences and to communicate relevant information to them.



B. Adapting form and content provides entry points into policy guidance

There is scope to reimagine the structure and format of policy documents to make them more relatable and actionable for product makers.

Product and policy do not speak the same language.

Long, detail-heavy documents with normative requirements can be difficult for product teams to parse, especially if they only need to consider a specific part of the policy guidance. Not all companies know how to factor this kind of policy into their workflow, particularly in startup contexts where there are limited resources to devote to this task.

This doesn't mean dispensing with text-based policy.

By adapting and augmenting existing policy content and formats, policymakers can enable product makers to better engage with guidance around AI explainability.

Aligning ethical commitments with policy guidance

What if

... policy included guidance for aligning company visions with high-level principles and ethical design approaches to responsible AI and transparency?

A product maker's ethical commitments can create a bridge between the worlds of product and policy.

When a company articulates a commitment to responsible and ethical AI, it is important they connect it to specific AI policy guidance. Signaling their readiness to embrace the principles and implement the requirements of a particular policy, such commitments allow product makers to cultivate better alignment between their aims and those of regulators.

In response to questions around their alignment with public policy or regulatory developments, a number of startups participating in the **People-Centric Approaches to AI Explainability** project, such as **X0PA AI**, cited national regulations as part of their commitment to ethical AI.

Commitments like this, grounded on actual policy guidance, have real impacts. They help teams and individuals within organizations implement the guidelines and measures provided by policymakers, setting them up to better incorporate responsible AI considerations into their operations and product development processes.

The opportunity for policymakers is to help product makers connect their ethical commitments to specific policy guidance. This can include providing tools that encourage companies to update and enhance their statements, in turn supporting product makers to operationalize responsible AI principles and driving industry maturity around AI explainability.

.

> "We are committed to ethical AI. We comply with Singapore's guidelines for responsible AI.
>
> – X0PA AI

Helping policy users find what they need

What if

... the principles of a policy were abstracted as a supplementary overlay, providing entry points for product makers to locate and navigate the specific parts of policy they need?

With a number of different audiences for a policy, policymakers need to ensure product makers can locate the provisions and guidance most relevant to them.

The needs of different developers and product teams represent a challenge for policymakers, both in terms of formulating policy and in helping them navigate the final document.

On the one hand, policy aimed at the AI industry as a whole may not necessarily provide detailed, meaningful guidance and orientation to individual product makers. On the other hand, a policy that attempted to address the specific contexts of different product makers and service providers would become unwieldy and difficult to navigate, besides being extremely difficult to create in the first place.

One way to make policy more accessible for product makers is to augment it with supplementary operational guidance. This could take the form of a series of prompts structured around the abstracted principles of the policy, similar to the ***Implementation and Self-Assessment Guide for Organisations (ISAGO)***. As a companion guide to the IMDA/PDPC's *Model AI Governance Framework*, the *ISAGO* poses a series of questions for consideration by product makers who procure and deploy AI solutions.

Drawing on the principles of self-directed learning, supplementary guidance can assist companies in assessing where they fit into a policy, locating and implementing the provisions most relevant to their context.

The findings of previous Open Loop policy prototyping programs in Europe, Singapore and Mexico support these approaches, highlighting the importance of guides, playbooks and toolkits for policy implementation. They provide vital entry points for product makers, helping them navigate and interpret a policy based on their particular situation.

Making policy clearer and more explicit

What if

... policy included calls to action, practical guidance on how to implement provisions and explanations on why specific actions should be performed?

Specify the Action

Write provisions in concise, action-oriented ways, focusing on the concrete tasks product makers should perform:

- *Ensure...*
- *Be transparent about...*
- *Ensure you have...*
- *State clearly...*

WHAT

Provide Instructions

Support required actions with guidance, supplemented by relevant how-to information as appropriate:

- Templates
- Tools
- Taxonomies
- Parameters
- Timeframes

HOW

Identify the Purpose

Connect product makers' actions to specific impacts and benefits, allowing them to understand what certain tasks are intended to achieve:

- *... to allow organizations to...*
- *... to support organizations to identify...*

WHY

Policymakers should be specific about what they want product teams to do and how they should put policy into practice.

The effective interpretation and implementation of a policy is fundamental to its impact. But even if a product maker is aligned with the overarching aims and is able to identify the provisions relevant to their context, actioning policy remains a critical challenge.

The risk is that without appropriate guidance, product makers may not effectively adopt or implement policy provisions and requirements.

In the policy prototyping workshops for this project, participants were asked to rewrite policy statements for different policy audiences. Their outputs, which had striking similarities, formed the basis for generating three principles for making AI explainability policy more accessible and actionable for product makers: *Specify the Action*, *Provide Instructions* and *Identify the Purpose*.

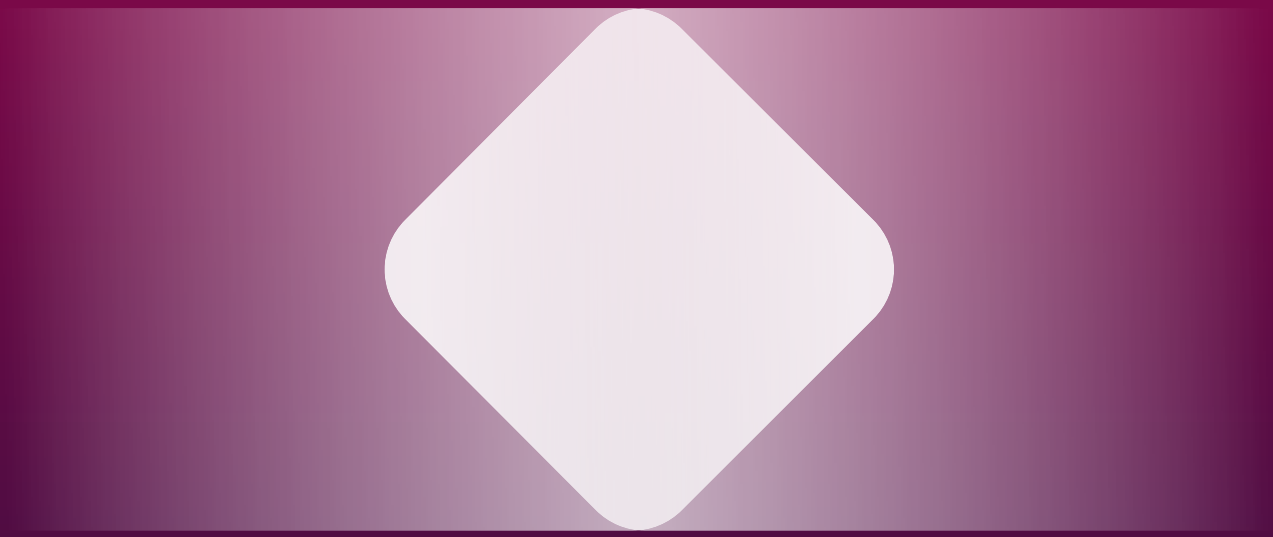
> *These observations and principles validate and complement previous findings, as documented in reports on Open Loop's policy prototyping programs in Europe (2021), Singapore (forthcoming 2022) and Mexico (forthcoming 2022).*

C. Collaboration drives better policy outcomes

Bringing together policy and product teams to create policy is key to realizing shared ambitions for people-centric explainability experiences.

It is clear that policy and product teams can offer each other valuable guidance. It is also clear that they are keen to learn from each other – policymakers want to be informed by product makers, while product makers are looking to policymakers to provide them with direction.

What needs to be understood more clearly is how these two fields can best collaborate to create actionable AI explainability policy and implement it effectively.



Aligning policy and product governance frameworks

What if

... policy were structured according to product governance principles? How can product governance frameworks incorporate more extensive considerations of societal needs?

Acknowledging the differences between them, there are opportunities for policy and product governance to complement and support each other.

Policy and product governance frameworks are not (and should not) be identical. Creating greater alignment between them, however, can strengthen the foundations for achieving common explainability goals.

What can policy learn from product governance, and what can product learn from policy?

Formulating policy in a way that aligns with product-making approaches and structures like **RAI's Explainability Framework** can help highlight shared governance aims, supporting companies to make sense of, and action, AI explainability guidance at a product level.

> "Aligning policy to a user flow is useful because it makes it clear how it will live in the real world
>
> – Workshop participant (industry expert)

Supporting the product design process

What if

... policy offered more detailed guidance for key moments in a user experience? Is there a way for policy elements to be incorporated into product design briefs?

To strengthen the connection between the product design process and the aims of a policy, policymakers need to determine the type and level of guidance to provide.

Transparency and explainability policies are by nature aimed at achieving collective benefits and protections. To deliver the best outcomes for the greatest number of people, they focus on the cumulative impacts of AI-powered products and services on society as a whole.

From a product design perspective, it's not always clear which mechanisms will most effectively achieve these outcomes. Traditional design briefs tend to focus less on wider societal concerns and more on people's needs in specific moments. They guide experiences that unfold step by step and screen by screen.

Policymakers therefore need to determine the type and level of guidance product teams require for the design decisions they make at each of these steps. Feedback from participants in the policy prototyping workshops identified potential opportunities to provide more detailed, granular guidance around key moments and milestones in a user experience, including:

- Clear parameters on the depth and extent of explainability required at different stages
- Tools explaining how to implement explainability information (e.g. a guide or playbook).

Another approach could involve developing AI transparency and explainability guidance for inclusion in design briefs, encouraging product makers to consider societal concerns in addition to the needs of individual users. This guidance should be informed and accompanied by an assessment of the types and levels of risks posed by AI systems. The type of AI, how it is used and the risks it raises will allow product makers determine the levels of transparency and control, privacy, and security messaging and controls they require.

Reimagining the relationship between policy and product

What if

... policymakers and product makers work together to explore what more generative spaces for collaboration might look like?

Realizing the aims of people-centric AI explainability will require ongoing collaboration between policymakers and product makers.

The policymakers and product makers participating in the Design Jam all indicated a desire to work more closely with each other, but acknowledged that collaboration is not without its challenges. If these opportunities are not commonplace, how might they come about?

One suggestion is that product makers' in-house legal teams could help bridge the gap between product and policy, facilitating better communication and understanding between these worlds.

Another approach involves shifting the preconceptions policymakers and product makers have of each other. In addition to recognizing product makers as policy users, policymakers might also regard these companies as policy co-creators.

Thinking in these terms might create more opportunities to engage product makers during policy development, review and implementation, utilizing product design insights to make policy more effective.

Product makers, for their part, want to be consulted. They want to contribute to the creation of definitions, standards and recommendations, ensuring policies don't introduce unmanageable requirements or have unintended consequences for industry.

Similarly, they want to have a voice when it comes to assessing the impacts of a policy and identifying areas for refinement. When a policy changes and evolves, what kinds of consultation might ensure they maintain alignment with the updated requirements?

For this kind of collaboration to be successful, both parties need to consider policy as a positive guiding force. Rather than regarding policymakers narrowly as enforcers, product makers should see them as active collaborators in providing new visions for addressing people's concerns and responsible AI considerations through product design decisions and processes. This could include involving policy experts at the beginning of the product exploration process, to provide input and actively think through unintended uses or outcomes.

Policymakers can support this lens shift by being clear on the value of AI explainability for product makers, cultivating a mature ecosystem through support and leadership that fosters both responsible technology development and economic growth. Increasingly, policymakers may seek to leverage design expertise and integrate design methods into the policy development process. Closer collaboration between policy and product makers will also enable policymakers to gain a better understanding of the impact of forthcoming regulation on the industry, namely in terms of technical feasibility, operational challenges and implementation costs.

Product makers and policymakers won't always agree – and nor should they. The aim of working together is to help both parties better understand each other's position and more fully explore the possibilities for AI explainability policy, thereby improving people's experiences of AI-powered products and services.

Next Steps

Contributing to cross-sector efforts to promote people-centric AI explainability

Three approaches to people-centric explainability

Highlighting the value of AI explainability for general product users.

As AI-driven technologies become more common in consumer-facing services, people increasingly want to understand how AI systems are affecting their experience of different products.

This report contributes to cross-sector efforts to address this need.

The draft **AI Explainability Framework** developed by Meta's Responsible AI team (RAI) provides guidance on the design of explainability experiences for AI-powered products for product makers across industry sectors.

The report also presents operational **Product Design Insights** and considerations for improving people's understanding of AI systems. Featuring examples of people-centric experiences of AI explainability in digital services, these learnings are intended to be used in conjunction with the Framework's guidance.

Lastly, the report details a series of **Public Policymaking Insights**, considerations and questions. These prompt policymakers to contemplate what might be involved in the development of policy guidance that promotes the kind of explainability experiences posited by the Framework and by the Product Design Insights.

Refined through multi-stakeholder consultation, this report and its findings are neither the beginning nor the end of the conversation around people-centric explainability. They point to a number of next steps and future opportunities, detailed in the following pages.

Future opportunities for the AI Explainability Framework

Meta is continuing to develop its tools for increasing transparency and control around AI systems.

With the publication of this report, RAI is promoting the Framework both at Meta and in the wider ecosystem, exploring opportunities to evolve it in response to cross-sectoral feedback.

> Meta ultimately aims to create an integrated transparency solution that translates model outputs into transparency features and controls for the people using its products

Key Framework opportunities

- **Further developing standard design patterns** to address commonly occurring explainability requirements
- **Conducting further research** on these design patterns, such as user testing
- **Providing samples of open-source code** with these design patterns, enhancing the ability of developers and practitioners to incorporate them into their products and services
- **Developing guidelines around the design of pathways** between different explainability dimensions for users
- **Creating awareness among industry**, including startups, leveraging the influence of incubators, accelerators and forums to facilitate the uptake of the Framework among developers and product teams

Potential wider opportunities for RAI

- **Developing values and principles** that may help product teams work through edge cases, adding to the stock of collective knowledge and giving back to the community
- **Developing product research playbooks** that may help product makers developing the next big innovations in AI to ensure they have the appropriate expertise in the room and access to responsible AI approaches to inform their design decision-making
- **Developing a service design approach to AI product making** that may help ML experts incorporate aspects of responsible AI seamlessly into their workflows, dashboards, outputs and practices
- **Codifying definitions and terms into a public glossary** that may reduce the cognitive load for product users, ML experts and external stakeholders

Future opportunities for product makers and policymakers

Improving the collective understanding of AI explainability techniques.

The findings of this project point to the value of ongoing research into the ways product makers can most effectively enable people's understanding of AI systems and processes.

Through cross-sector collaboration, product makers and policymakers can undertake further research into explainability techniques and explore methods for assessing the degree to which people comprehend AI systems.

Together with this report, ongoing research can help policymakers facilitate the creation of people-centric explainability and trustworthy AI experiences. In consultation with industry, they can improve the impact of these policies by incorporating strategies that make them more actionable for product makers.

Key cross-sector opportunities

- **Exploring the evolving nature of people's trust** in technology and their need for policy and product mitigations, including the recognition of tipping points where particular conventions have been normalized within society and mitigations may no longer be required
- **Mapping code frameworks and snippets** to policy requirements, developing engineering tools that help generate and assess the levels of transparency and control required for different use cases and levels of risk
- **Developing international standards** for AI transparency and explainability
- **Developing explainability assessments** to understand whether appropriate information has been provided to people whose experiences have been affected by AI systems
- **Utilizing experimental governance methods** such as policy prototyping and sandbox programs to further explore explainability techniques and user controls, testing different use cases and product solutions against practical guidance such as the Product Design Insights detailed in this report
- **Exploring the emerging need** for greater AI accountability, transparency and documentation for the benefit of expert audiences such as regulators, legislators, academics and others tasked with holding AI to account
- **Engaging with academic institutes** to ensure regulations are considered as part of the curriculum in computer science programs

Appendices & References

Appendix A

Who are TTC Labs, Open Loop and IMDA/PDPC?

About TTC Labs

Initiated and supported by Meta, **TTC Labs** drives collaboration between policymakers, privacy experts and technologists through design thinking. We build trust, and we advocate for transparency and control, for Meta platforms and for digital services around the world. Our aim is to focus on what people across the globe need, want and require from technology. We need to keep working together for a scalable approach to building trust, transparency and control into data-driven products and services. Our vision is to create meaningful relationships between people and data that are sustainable and equitable for all.

To date, TTC Labs has brought together more than 300 industry and design companies as well as 200 policy, academic and civil society organizations globally to tackle shared challenges. These challenges include notification and consent, explaining data concepts to different audiences, algorithmic transparency, privacy and digital literacy, augmented and virtual reality, and designing for young people.

TTC Labs creates materials that anyone can use, adapt and replicate. We publish design solutions and reports that synthesize learnings and insights from co-creation workshops called Design Jams, enabling the wider community to collaborate on shared challenges. We develop interactive exercises and visual explainers to support understanding and exploration. And we share our open-source **Toolkit** to support designers and practitioners.

Together with our network of design partners, including expert agencies operating in key regions around the world, we actively foster collaboration and innovation in order to speculate on potential solutions and future-facing approaches to driving digital trust, transparency and control.

About Open Loop

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

Open Loop creates a robust collaborative feedback loop (an 'open loop') of practical learnings between the people who make policy and those who are required to implement it. As part of Open Loop, policymakers work with a vibrant community of technology companies to build sound and operational governance frameworks and discuss regulatory best practices. Participating companies leverage training, tutorials, toolkits, mentorship and technical assistance while sharing practical insights and working directly with policymakers to inform new regulatory solutions. Open Loop takes an experimental, interactive approach to policy prototyping, informed by and aligned with product development processes: alpha phases to research and test different regulatory pathways; and beta phases to iterate and refine these frameworks before sharing them more broadly.

About the IMDA and PDPC

The Singapore **Infocomm Media Development Authority (IMDA)** and **Personal Data Protection Commission (PDPC)** develop and regulate the converging infocomm and media sectors in Singapore in a holistic way, creating a dynamic and exciting sector filled with opportunities for growth, through an emphasis on talent, research, innovation and enterprise. Singapore sees AI as an important and emerging technology for the digital economy. The IMDA and PDPC have released a suite of AI governance initiatives to help organizations deploy responsible AI and build consumer trust. These include the *Model AI Governance Framework* and the *Implementation and Self-Assessment Guide for Organisations (ISAGO)*, Volumes 1 and 2: Compendium of Use Cases.

Appendix B

What is product and policy prototyping?

This report presents findings from collaborative workshops in the form of future-facing insights and considerations for product makers and policymakers. Our process for these fast-paced participatory workshops followed two approaches:

Product prototyping

to co-design AI explainability solutions for startup products through Design Jam workshops facilitated by TTC Labs.

Policy prototyping

to test AI explainability governance frameworks and derive evidence-based insights to improve policymaking processes through workshops facilitated by Open Loop.

The product prototyping Design Jam sessions and the policy prototyping workshops both brought together stakeholders from government, academia, civil society and industry to rapidly prototype ideas, design patterns and insights.

Developed by TTC Labs, Design Jams are interactive co-creation workshops. These output-oriented sessions bring experts together to experiment with different methods and interfaces that put people at the center of how we design for trust, transparency and control in the digital space. During Design Jams, product makers work alongside policymakers, academics and members of civil society organizations to solve both real-world design problems and proxies of real-world design problems. Through hands-on product prototyping on real and fictional digital products, multidisciplinary teams co-create innovative solutions to challenges around trust, transparency and control.

TTC Labs Design Jams and Open Loop workshops both create an environment that fosters collaboration between different stakeholders for product and policy prototyping. No single group has all the answers, so these sessions provide a real opportunity to experiment in a judgment-free way. This co-creation process demonstrates the importance of foregrounding people's needs both for the design of data-driven experiences and in data policymaking, informing complementary product and policy pathways.

In September and October 2021, TTC Labs – in collaboration with Asia-Pacific design agencies Craig Walker (Australia, Singapore) and Wunderman Thompson (Indonesia) – facilitated four virtual product prototyping Design Jam sessions, each between two and four hours in length.

The sessions focused on:

- **Defining personas and product challenges** – Multidisciplinary startup teams identified and built out a persona (a fictional representation of a real user affected by AI explainability), which they used to articulate a question that framed their AI explainability design challenge
- **Ideating and prototyping design patterns** – The same teams sketched and discussed ideas that responded to their AI explainability design challenge, collectively refining their ideas into a design pattern or user interface (UI) which they developed and pitched as a solution.

In October 2021, Open Loop – in collaboration with University of California, Davis – facilitated two virtual policy prototyping sessions, each between two and three hours in length. These involved rapid exploration and experimentation to build an evidence base, generating and iterating on insights that drew on personas, user journeys, frameworks and observations from product prototyping as well as policy guidance.

Finally, this report draws on design patterns co-created for real and fictional products at previous Design Jams held in Singapore (May 2019) as well as prototypes co-created on fictional AI-driven apps at workshops held in Washington DC (October 2019) and Amsterdam (December 2019).

Explore the TTC Labs Design Jam methods and resources in detail, including our Personas for AI explainability, in our open-source Toolkit at **toolkit.ttclabs.net**. In particular, you can **create your own personas** or use our ready-made **personas for AI explainability**.

Appendix C

Project observations on the AI Explainability Framework

The **People-Centric Approaches to AI Explainability** project provided Meta's Responsible AI team (RAI) with a collaborative, practical setting to test the **AI Explainability Framework** on AI-powered products.

Participating startups were encouraged to use the Framework as a prompt and a reference throughout the workshops, and the prototype solutions they co-created were analyzed and interpreted through the lens of the Framework and its dimensions of explainability.

Our observations from this process validated key aspects of the Framework and RAI's internal research, documented in this Appendix, and identified some important future opportunities for this work, as detailed in the **Next Steps** section of this report.

Mapping prototypes to the Framework

The AI explainability prototypes developed by the multidisciplinary Design Jam teams map neatly to the Framework. In each of the prototypes we can identify the Framework's explainability dimensions, either implicitly or explicitly, providing people with information relevant to their needs and expectations in the context of the respective startup product.

These correspondences reinforce the four-level structure of the Framework. Each of the Framework's dimensions can be identified in the prototypes as representing a unique type of information, supporting the division of AI explainability into these four levels.

	Betterhalf.ai	MyAlice	The Newsroom	X0PA AI	Zupervise
1. AI Awareness	●	●	●	●	●
2. AI Outcome Explainability	●	●	●	●	●
3. AI Product Explainability	●	●	●	●	●
4. AI Model Explainability					●

Explainability dimensions that feature either explicitly or implicitly in the respective startup prototypes.

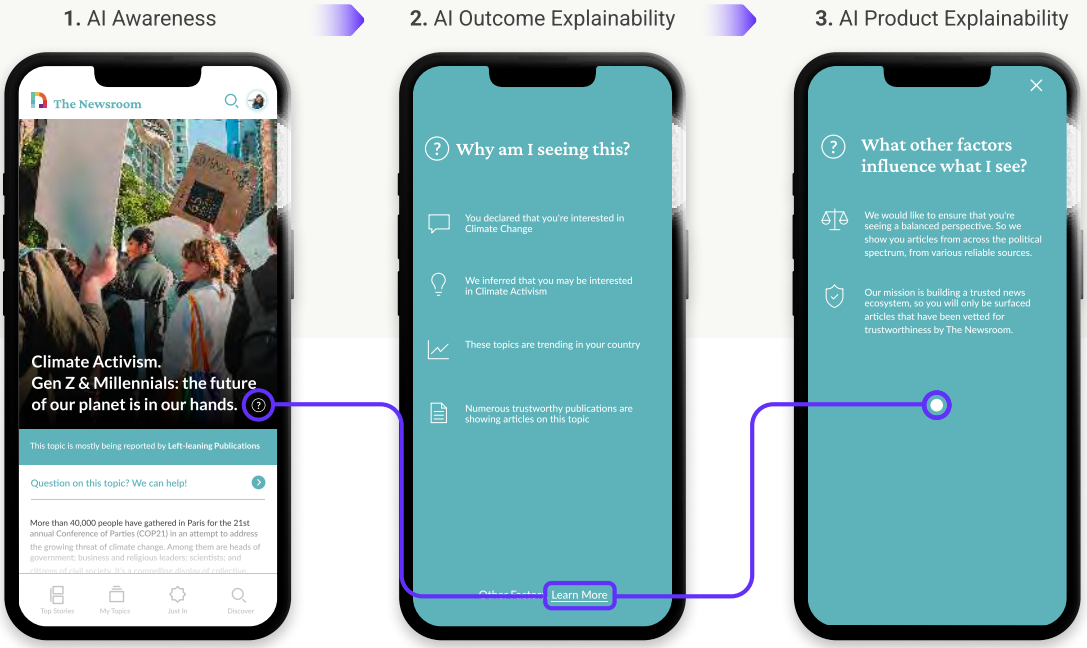
● Implicitly ● Explicitly

Pathways between explainability dimensions

Explainability isn't just about driving awareness and understanding of AI – it should also provide information and controls around privacy and security where applicable. In this way, transparency and control can become vehicles for other aspects of responsible AI practices that should be afforded to people using AI-enabled products.

The startup prototypes provide instructive examples for building pathways between different dimensions of explainability, allowing people to navigate between different levels of information and control. None of the prototypes map to a single level of the Framework, with each incorporating two or more dimensions on the same screen or connected through a user flow. In this way, the AI explainability solutions illustrate the interrelated nature of the Framework dimensions.

The Newsroom's prototype shows clear pathways between different explainability dimensions



Different stages of the user journey

As the multidisciplinary Design Jam teams moved from framing and exploring their AI explainability challenges to developing screen flows and prototype solutions, their design patterns began to align with different stages in the user journey. These stages translate to distinct explainability touchpoints – upfront, in context and on demand – as detailed further under the Product Design Insight **People need different information at different stages**.

The prototype solutions validate previous research and thinking in identifying and categorizing these touchpoints. They also validate RAI's assumptions that different touchpoints require different types of explainability, and provide an indication of the specific dimensions that align with each touchpoint.

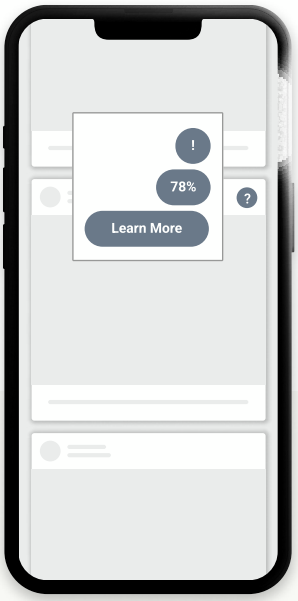
	Upfront	In Context	On Demand
1. AI Awareness	●	●	●
2. AI Outcome Explainability		●	
3. AI Product Explainability	●	●	●
4. AI Model Explainability			●

Standard Design Patterns

Prior to this project, RAI's research had identified the opportunity to incorporate example design patterns into the Framework. The startup prototypes confirmed this opportunity.

The design patterns included as part of the **AI Explainability Framework** were developed in reference to the use cases explored during the Design Jam. Each template reflects one or more of the explainability experiences prototyped by the multidisciplinary teams.

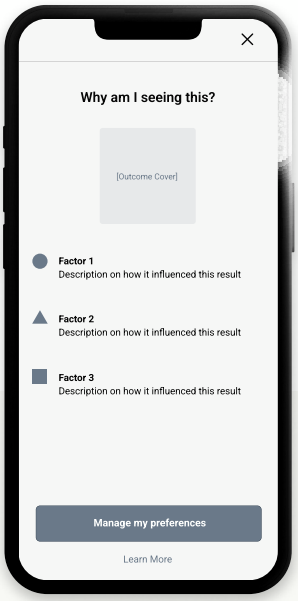
These templates demonstrate standard approaches to surfacing different explainability dimensions within common user experiences.



Awareness Sticker

1. AI Awareness

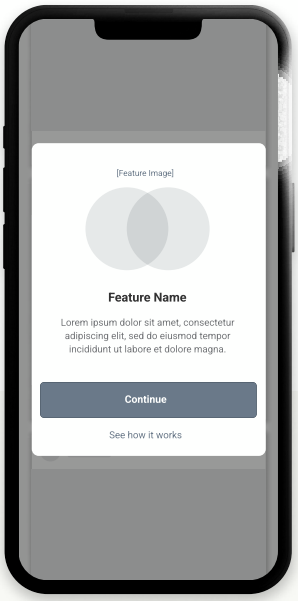
The Newsroom, Betterhalf.ai



Outcome Explanation

2. AI Outcome Explainability

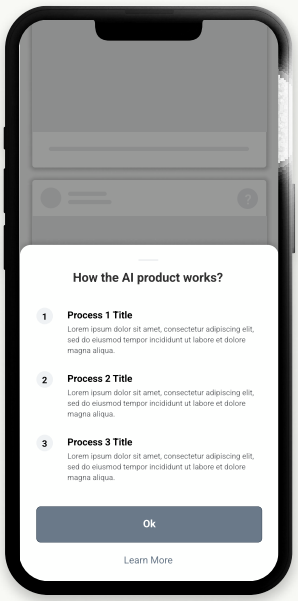
The Newsroom



Product Onboarding

1. AI Awareness

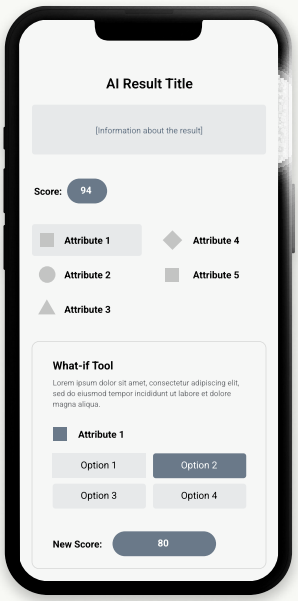
Betterhalf.ai



Product Explanation

3. AI Product Explainability

The Newsroom, Betterhalf.ai, Zupervise



What-if Explanation

2. AI Outcome Explainability

X0PA

These consolidated patterns offer product makers suggested guidance on the design of commonly occurring explainability requirements in their products. They promote the provision of explainability in a consistent manner across product ecosystems, helping to build familiarity, understanding and agency for the people who use these products, in turn improving the societal outcomes for AI-driven technology over time.

References

Some resources you might find useful

Ada Lovelace Institute. *Regulate to innovate: A Route to Regulation that Reflects the Ambition of the UK AI Strategy*. Report. (2021)

Ananny, M., and Crawford, K. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." (2016)

Business at OECD (BIAC). *Regulatory Sandboxes for Privacy Analytical Report*. (2020)

Google. *People + AI Guidebook.* (2019)

Infocomm Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC). *Model Artificial Intelligence Governance Framework.* 2nd edition. (2020)

IMDA, PDPC and World Economic Forum (WEF). *Implementation and Self-Assessment Guide for Organisations (ISAGO).* (2020)

Information Commissioner's Office (ICO) and Alan Turing Institute. *Project explAIn.* Interim Report. (2018)

INSEAD. "Implementing AI Principles: Frameworks, Processes, and Tools." (2020)

John, Peter. *Analyzing Public Policy*. London, Routledge. (2012)

Lucic, Ana, et al. "A Multistakeholder Approach Towards Evaluating AI Transparency Mechanisms." arXiv preprint arXiv:2103.14976. (2021)

Maiorana, T. "The Failure of Prototyping: A Call for a New Definition." Proceedings of Relating Systems Thinking and Design Symposium. *Relating Systems and Design* 10, TU Delft, Netherlands. (2021)

Meta. "Facebook's Five Pillars of Responsible AI." (2021)

Meta. "Instagram Feed Ranking System Card." (2022)

Meta. "Privacy Progress Update." (2021)

Meta. "System Cards, a new resource for understanding how AI systems work." (2021)

OECD. "Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449." (2021)

OECD. "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems." *OECD Digital Economy Papers*, No. 312. (2021)

Open Loop. *AI Impact Assessment: A Policy Prototyping Experiment.* Report. (2021)

Open Loop. *AI Transparency and Explainability: A Policy Prototyping Experiment.* Report. (2022, forthcoming)

Phillips, P., et al. "Four Principles of Explainable Artificial Intelligence." National Institute of Standards and Technology (NIST). (2021)

Primmer, N. and Andrade, N. "Businesses are applying the OECD AI Principles. How is it going?" *OECD.AI*. (2021)

Stanford University. *Artificial Intelligence Index Report 2022.* (2022)

TTC Labs. *People-centric Approaches to Notice, Consent, and Disclosure.* Report. (2020)

TTC Labs. "Making Sense of Data Disclosures: Leveraging Context in Design." Article and Visual Explainer. (2020)

Vilone, G, and Longo, L. "Explainable Artificial Intelligence: A Systematic Review." arXiv preprint arXiv: 2006.00093. (2020)

Warner, R, and Sloan, R. "Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables." *Criminal Justice Ethics* 40. (2021)