AI Transparency in Practice

Builders on what works — and what doesn't. Mozilla's research on AI transparency, with practical advice from Thoughtworks

By Ramak Molavi Vasse'i , Jesse McCrosky

Research lead: Ramak Molavi Vasse'i

This report presents the interdisciplinary research led and conducted by Ramak Molavi Vasse'i (Mozilla's Insights Team). Jesse McCrosky (Thoughtworks) co-authored the report to add some practical perspective based on his desk research and industry experience in responsible AI.

March 2023

Table of Contents

Executive Summary

Transparency is at the heart of responsible AI. In this study, we explore the concept of meaningful AI transparency, which aims to provide useful and actionable information tailored to the literacy and needs of specific stakeholders. We survey current approaches, assess their limitations, and chart out how meaningful transparency might be achieved.

The study was conducted in light of upcoming regulation, such as the AI regulation in Europe's AI Act,[1] the Federal Trade Commission's increased attention to AI,[2] and the US AI regulation on the horizon.[3]

Based on surveys and interviews with 59 builders with transparency expertise from a range of organizations, the report examines the current state of AI transparency and the challenges it faces.

Findings include low motivation and incentives for transparency, low confidence in existing explainability tools, difficulties with providing meaningful information, and a lack of focus on social and environmental transparency. The report highlights the need for greater awareness of and emphasis on AI transparency, and provides practical guidance for effective transparency design.

In the absence of adequate ex-post explanation solutions, we encourage builders to consider using interpretable models rather than black-box solutions for applications in which traceability is a design requirement. We aim to build a community around best practices and solutions and raise awareness of transparency frameworks and methods.

[1] European AI Act Proposal, April 21, 2021
[2] Elisa Jillson, "Aiming for truth, fairness, and equity in your company's use of AI", April 19, 2021
[3] Alex Engler, "The EU and U.S. are starting to align on AI regulation" February 1, 2022

Key Findings

- The focus of builders is primarily on system accuracy and debugging, rather than helping end users and impacted people understand algorithmic decisions.
- AI transparency is rarely prioritized by the leadership of respondents' organizations, partly due to a lack of pressure to comply with the legislation.
- While there is active research around AI explainability (XAI) tools, there are fewer examples of effective deployment and use of such tools, and little confidence in their effectiveness.
- Apart from information on data bias, there is little work on sharing information on system design, metrics, or wider impacts on individuals and society. Builders generally do not employ criteria established for social and environmental transparency, nor do they consider unintended consequences.
- Providing appropriate explanations to various stakeholders poses a challenge for developers. There is a noticeable discrepancy between the information survey respondents currently provide and the information they would find useful and recommend.

1. Meaningful AI Transparency

Done right, transparency is a means to an end. The right information - delivered in the right way - is the key. Meaningful transparency aims to ensure that each stakeholder receives an explanation that is both adequate and understandable, by providing useful and actionable information to enable them to make informed decisions. Information provided through meaningful transparency can help stakeholders object, opt out, assess fairness, or identify the responsible entities behind the AI.

Transparency done wrong can be overwhelming, leading to transparency fatigue, a kind of exhaustion and cynicism that can be disempowering. Similarly, transparency delivered in the wrong way can create the illusion of meaningful control, potentially constituting a deceptive design. This reflects the tension between transparency created simply for the sake of more transparency versus transparency that is designed to actually create more trust and accountability. A delicate balance has to be met to preserve human control and responsibility and to navigate these dilemmas.

Researchers at the University of Groningen stated that, “humans require a narrative form of explanation which opposes the binary nature of AI systems’ outputs.”[4]
Again other researchers note that “most citizens would not trust any AI system if they were simply told, 'We can't explain how it works, but it's really safe.'”[5]

Substantial work is being done to develop and evaluate AI explanation tools, or XAI. As we show in this report, trust in such tools is generally low. Such tools are oriented towards creating AI transparency and accountability, and they can be somewhat effective, but, as we discuss in Section 5, in cases where interpretability and traceability is a requirement, it is preferable to create a model that is interpretable by design, rather than trying to create explanations post-hoc.

As well, AI transparency is much broader than just technology focussed explainability/interpretability. Other methods can be more important for creating meaningful transparency. This could look like communicating the risks and limitations of a system to users and consumers, providing information about the person behind the algorithmic decision who can be held accountable, sharing how stakeholders can initiate communication about decisions that affect them, distributing documentation that breaks down the specifications used for the decision making, or offering post-decision verification letters.

[4] Ida Varošanec, “On the path to the future: mapping the notion of transparency in the EU regulatory framework for AI” (2022)

[5] Reed et al., “Non-Asimov Explanations Regulating AI Through Transparency” (2021)

Discussions about AI transparency often focus on explainability, which is seen as a central element.[6] However, **AI transparency and explainability are not the same thing.** Explainability is only a subset of AI transparency, as shown in the figure below. The European Parliament calls for explainability in addition to transparency in its 2020 recommendations[7] on a framework for ethical aspects of AI, robotics, and related technologies. Taking a broader view on AI transparency, our figure below shows elements that may qualify as AI transparency beyond explainability, such as the purpose of use, the metrics of the AI, the provenance of the data, information about its potential societal and environmental impacts, as well as the clear attribution of responsibility. Social transparency involves considering and disclosing the social, organizational, and cultural context of the use of an AI system.[8]

This includes providing information about the factors that may have influenced the development of the technology and its use in society, as well as considering how the technology may affect different social groups and communities.

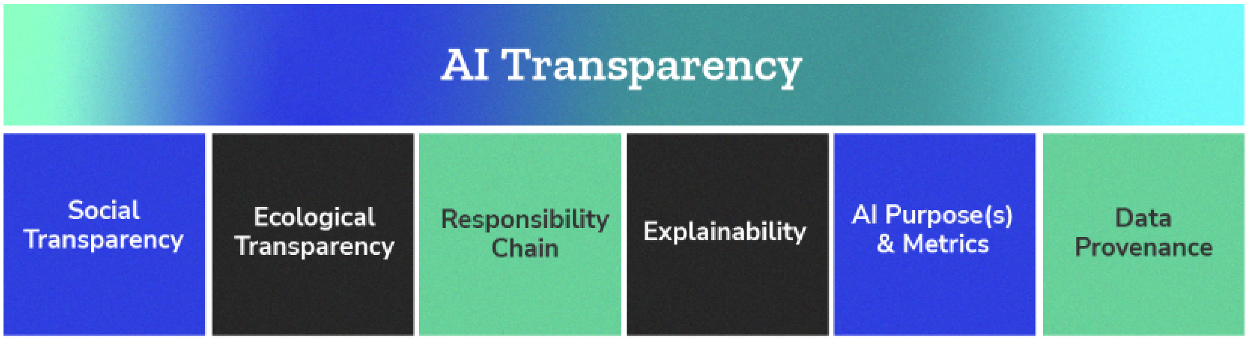


Figure: Information that may be considered elements of AI transparency beyond — and including — explainability.

Many of these elements can be understood from a systems design perspective. In his paper "The fallacy of inscrutability"[9], Joshua Kroll writes about taking a design-oriented perspective on understanding AI:

> *Systems can be understood in terms of their design goals and the mechanisms of their construction and optimization. Additionally, systems can also be understood in terms of their inputs and outputs and the outcomes that result from their application in a particular context.*

[6] Markus et al., "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies" (2021)
[7] EU Framework of ethical aspects of artificial intelligence, robotics and related technologies, (October 20, 2020)
[8] Ehsan et al., "Expanding Explainability: Towards Social Transparency in AI systems" (2021)
[9] Joshua A. Kroll, "The fallacy of inscrutability" (2018)

So, simply put, in order to understand an AI system, we need to understand what it is designed to do. In the case of an AI system, there are certain characteristics that are especially critical to understanding the system design. These include understanding the data on which the system was trained, the features of the data, and the objective of the system.

There is a need to develop a multidisciplinary understanding of transparency in AI. We have been listening and learning from the community of builders to better understand their current approaches and the constraints and opportunities they face.

2. Transparency Stakeholders & their Needs

Different stakeholders need different information tailored to their transparency needs. Regulators and end-users have different needs. And the information needed to identify potential improvements in a system is not the same as the information needed to establish accountability.

It is clear that care must be taken to understand the transparency needs of each stakeholder. A proper understanding of these needs should, of course, be gathered in collaboration with representative stakeholders. In general, however, stakeholders include:

Builders
These are the developers and deployers who are directly involved in the design and creation of AI systems, or at a later stage in the AI lifecycle. They are data scientists, system architects, machine learning or data engineers, (highly involved) product owners and UX designers. They may also be whistleblowers who help create more transparency in the system. It's important to note that with the advent of general purpose AI systems like the GPT models, there is an increasing bifurcation between builders and deployers, where one organization might build a model, while a completely different organization might deploy that model in a product.

Clients/Expert-users
Those who use and interact with algorithmic systems, mostly in their work, and who are often paying customers. These people operate the AI developed by others in an expert domain (e.g. hospitals or an organization) or are domain experts, such as the doctor who uses AI to find a cancer pattern and the HR professional who uses a recommender system to find the top five candidates from hundreds of CVs. It can also include

advertisers who work with AI recommendation engines to promote their content.

End Users

End users are the recipients or consumers of a product or service. They can be the buyers (an eBay user who buys an item directly through the platform) or content generators (an Instagram user who posts Reels) using the algorithms to create or share their own content. Those who pay to include advertising on a platform are generally classified as clients, while those posting organic content for free are generally end users.

Subjects/Impacted

Subjects are individuals or organizations (usually small and medium-sized enterprises or SMEs) that are affected by an algorithmic decision or prediction. Subjects can be consumers affected by differential pricing, delivery services drivers whose pay is algorithmically determined, or patients whose doctors have used AI for diagnosis. Anyone can be affected by AI-based decisions, for example, in the case of an automated decision about welfare fraud or justice systems using AI to predict the risk of recidivism.

Society/Public

Sometimes, a risk may not impact individuals, but instead represent a broader social risk. For example, if someone benefits from the fact that their health insurance company offers a more favorable rate because of their extremely healthy lifestyle — in deviation from the principle of solidarity — this can have negative consequences for other insured persons who cannot afford such a healthy lifestyle and whose premiums then increase. In other words, what is good for an individual is not necessarily good for society.

NGOs, Watchdogs

Due to the lack of sufficient AI Transparency, nongovernmental organizations (NGOs) and Watchdogs play a crucial role in dismantling the needed information. Together with the independent AI research community and other academics and journalists, they investigate and liberate information and explain its significance to the public and affected communities. This enables actions and contestability. They also challenge information and expose cases of misinformation or transparency washing (providing incomplete, wrong, misleading information just to meet transparency demands).

Regulators/auditors

These include policy makers, enforcement bodies, standard setting bodies, and independent 3rd party auditors. Comprehensive and complete information is needed in order to develop the right form of governance or to set standards and, above all, to verify and enforce their implementation. It is also necessary to monitor the effectiveness of legislation. External, 3rd party audits are also only possible if access to the necessary information is available. The level of transparency and the depth or technicity of information is dependent on the specific use within that range of responsibilities.

3. Our Participants & their Transparency Work

We had 52 survey participants and 7 interviewees. 19 of the respondents completed the entire survey with its 48 questions. The findings presented in this report are accompanied by information on the number of responses.

For the survey, we reached out to a niche group of builders with experience in bringing transparency to AI. All of them are working on AI transparency issues within their organizations. 5 Developers; 9 Software engineers; 9 ML engineers; 9 data scientists; 20 “others” (includes product designers, product managers, QA testers, software and hardware support engineers, internal algorithm auditors, and machine learning researchers).
You can find more information about our research methodology below.

a. Intended domains of use

As reflected in the chart below, most participants in this research are building AI systems intended for use in government, public service, health care, and finance.

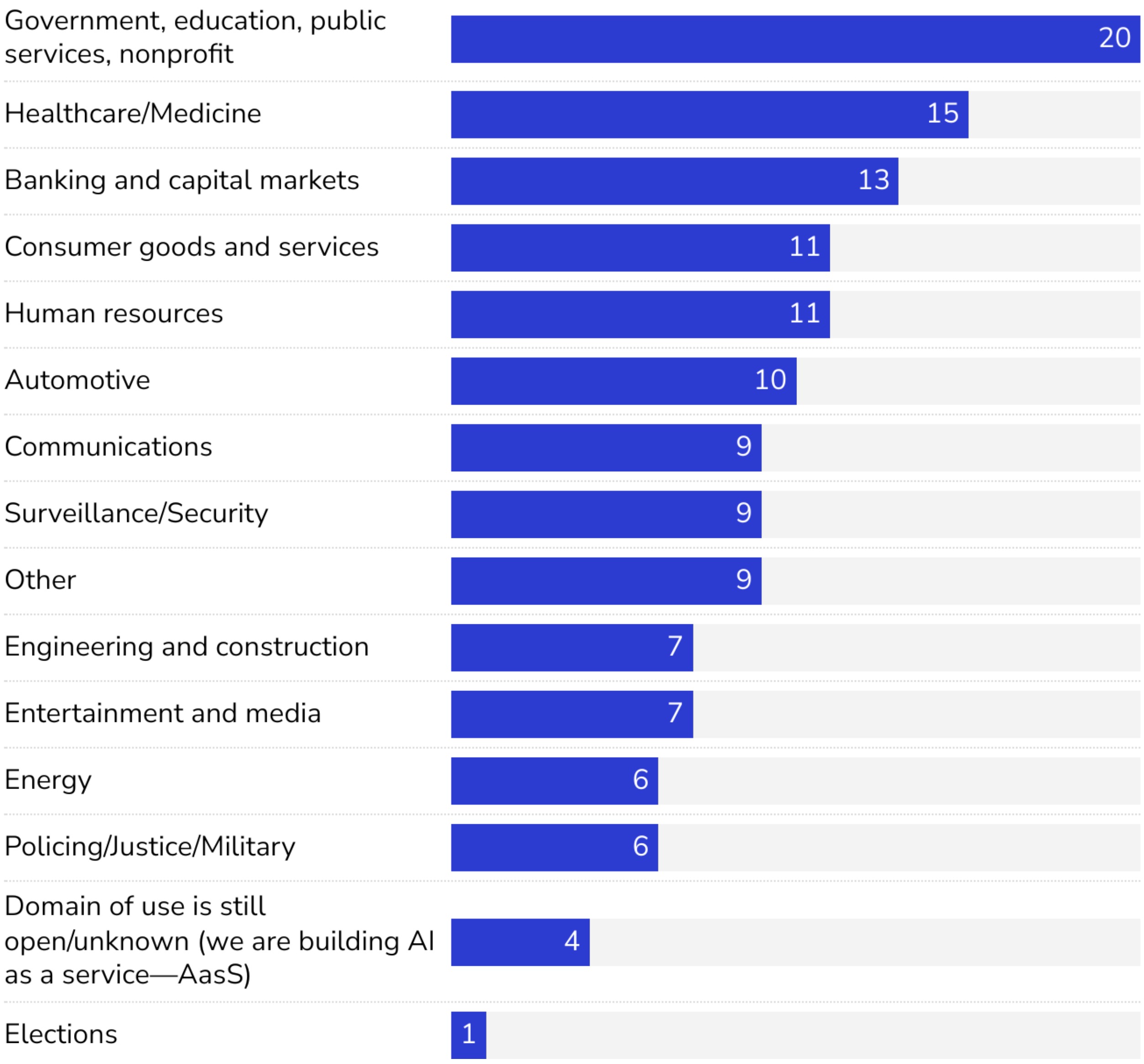


Figure: Intended domain of use for AI developed by research participants. Total responses: 51, multiple choice.

b. AI applications

The graph below shows the kind of AI systems participants work with. The majority of our participants who are working on AI transparency are working on natural language processing or forecasting/predictions. Note that respondents were able to select multiple system types.

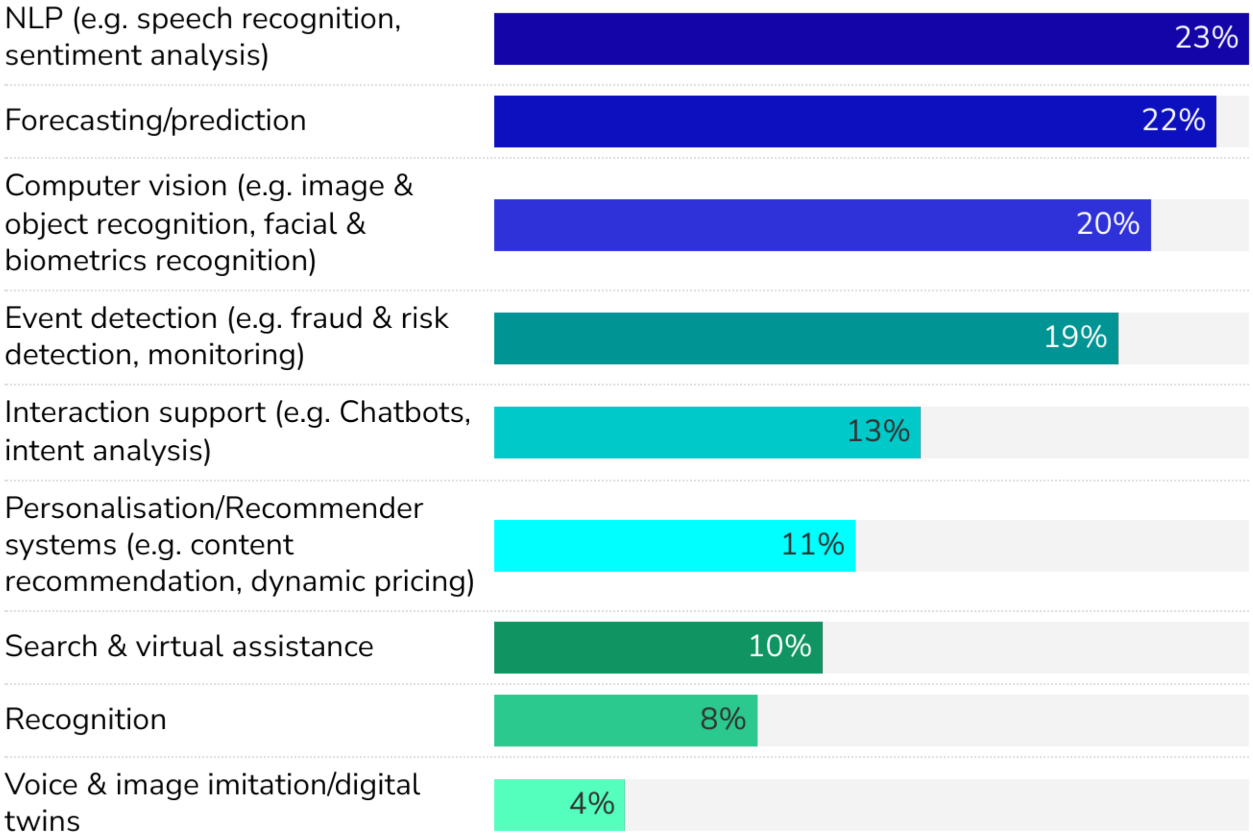


Figure: Types of AI systems used by 52 participants, multiple choice.

c. Knowledge about the provenance of the data used

Understanding the source of data – and how it was collected and processed – is key for obtaining trustworthy output. Bias can enter data through the process in which the data was collected, or even in the process through which it was generated, in which case the data reflects biases that already exist in society. As an example, if data is collected on what sorts of people get a particular disease, but the cases are measured by hospital diagnosis, we may have bias in the data-generating process from the fact that certain communities may be less likely to seek healthcare for the condition or get a prescription. And we may have bias from data-collection if data is only collected from certain healthcare facilities, which might have tended to focus on serving only particular types of people.

There are many sorts of biases. For example, if the data was collected through online surveys, it may be biased towards individuals with internet access, while data collected through door-to-door surveys would have a different bias. Additionally, it is important to consider the scope of the data, whether it's a global dataset or limited to a specific region or demographic. Understanding these factors can help identify potential biases that an AI system might perpetuate or amplify.

If data is collected through a survey in English, we are likely to miss those with lower English-language skills. Many other forms of bias are possible. It is essential to understand misalignment between the population over which data was collected and the population over which the model will be applied. For example, if a model is trained on data from contractors at an agency in one geographic region, it's unreasonable to expect that the model will accurately model people in other regions, or even the general population of that region.

It is also important to understand the features in the data, or what information the data contains – does it record home addresses? Web histories? It is essential to understand not only the features that were explicitly included in the dataset but also those that may be inferred. For example, if a model uses a person's home location as a feature, it may also be effectively using their ethnicity and socioeconomic status as features; this is because, at least in some parts of the world, knowing what neighborhood someone lives in tells you a lot about how wealthy that person is. Various studies[10] have shown that data such as interests/likes and location is highly predictive of diverse socioeconomic factors including: sexual orientation, political orientation, membership in certain ethnic groups, gender, age, education, and marital status. This has important design implications: the system may not be intended to make decisions based on ethnicity, but the features selected result in a design that is misaligned with this intent.

An important tension to consider is the potential trade-off between privacy and fairness. In many cases, collecting more (personal) data can reduce bias – an optional survey will always be biased by the types of people who choose to respond, whereas a mandatory survey such as a census will be much more representative of the population it covers. Similarly, it is difficult to impossible to reduce gender bias in a model if data on gender are not collected. In both cases, whether collecting data on more people or collecting more information about each person, more data can lead to fairer models.

[10] For a list of related research, see https://www.researchgate.net/publication/281717718_You_Are_Where_You_Go_Inferring_Demographic_Attributes_from_Location_Check-ins and https://www.pnas.org/content/110/15/5802

It is clear that choosing data – how and what is collected, from whom, and how it is processed – is a critical aspect of the design of an AI system. For this reason, we asked if the builders who took our survey understand the source of the data they use. Eighty-six percent of the 22 respondents to that question stated that they know the provenance of the training and test data of their systems. Only 14% answered "no."

Of the 55 people who responded to an earlier question on data, 60% said that they use proprietary data. About one-third of the participants use third-party data to train their models, and some of this data comes from open source datasets. The open source datasets used by our participants range from "whatever is available" to a very specific use depending on the field, including the following sources:

- **English Wikipedia**, datasets for natural language processing
- **DebateSum**, a large-scale argument mining and summarization dataset
- **MNIST**, a large database of handwritten digits that is commonly used for training various image processing systems
- **ImageNet**, a large visual database designed for use in visual object recognition software research
- **Mozilla's Common Voice Platform**, datasets for natural language processing
- **River Level Data (Scottish Environmental Protection Agency)**, a set that provides water level data from 335 sites
- **VoxCeleb**, an audio-visual dataset of short clips of human speech, extracted from interview videos uploaded to YouTube; used for speaker identification, speech separation, and talking face synthesis

d. Metrics alignment and system design

The true purpose of an AI system is encoded in the metrics it uses. Thus, understanding the purpose of a system requires understanding the metrics that the system is predicting, or for which it is optimizing. Every modern AI system has some sort of metric being optimized for under the hood and understanding the metric can shed light on the true, but potentially unstated, purpose of the system. For example, an AI system to set prices might be advertised as helping find prices that reflect market conditions, when it's really optimization to maximize profits for property owners.[11]

[11] Heather Vogell, "Rent Going Up? One Company's Algorithm Could Be Why." (2022)

Similarly, the YouTube claims that their recommendation algorithm will ”help you discover more of the videos that you love”[12]. In reality, we don't know precisely what combination of metrics are used, whether they are appropriate for the purpose, or how success is quantified. There is, however, overwhelming empirical evidence[13] that people frequently encounter content they find disturbing and not in line with preferences they feel they have communicated to the platform.

The purpose of a system may be expressed through the metric the model optimizes for, through what the model attempts to predict, or some combination of these and other factors. Generally, it is important to align between what the system is designed to do and what stakeholders are told the system is doing, as well as any constraints or limitations that may affect its performance. As in our example, a video recommender system's objective may be to show the video that the user is most likely to watch, but this may not align with the user's actual preferences or interests. There is a critical difference between showing the video the user will love compared to that which the user is most likely to watch. It's important to understand these potential mismatches and consider how they may impact the system's performance and the user's experience.

That’s why we asked our participants about metrics. The majority (92%) of the 24 respondents to this question said that they understand what metric the AI they are working on is optimized for. However, in the following question, with 25 responses, only 68% felt that the metric and stated goal of the AI system match, while the remaining 32% felt they differ or otherwise doubt that the metric is appropriate. In one case, the respondents felt that the metrics would not always reflect the best outcome for the end user, even at peak performance. Another reported that the general metric didn’t take into account vulnerable groups of people, leaving room for biases and drifts.

In interviews, we heard that selecting the right metrics in alignment with the design goals and intended purpose of a system requires a deep understanding of the model behavior on the builder’s side; understanding the problem in the specific context and matching the metric accordingly is key. However, making these decisions well can require deep multidisciplinary expertise in the technical aspects of the system, as well as business objectives and even psychology. Further, it is very easy to select metrics that “seem good enough”. For example, it is commonly assumed that if a user clicks on an ad, it means that the user in question finds the ad useful. However as per Goodhart’s law[14], aggressively optimizing for clicks can amplify the reach of ads that are “designed to attract attention and to entice users being typically deceptive, sensationalized, or

[12] Youtube’s description of the product feature “Recommended videos” (February 2023)
[13] Mozilla’s “YouTube Regrets” Report (2021)
[14] Wikipedia on “Goodhart's law”

otherwise misleading" known as "clickbait"[15] content. This is largely because it is generally easier to incentivize clicks using deceptive design than it is to create ads with true user value. Often, developers may not consider the full consequences of the metrics they select. This observation is exacerbated when applied to cases where the target use is not yet fixed, is flexible, or may change at a later date. In our survey, around 8% of AI models are stated to be built for multiple use. These cases deserve special attention and can be potentially critical due to decreased observability when the AI is repurposed after deployment and the increased potential for not closing the information loop back to the developer team.

It's important to note that the emergence of so-called "general purpose AI", such as ChatGPT can further muddy the waters. Transparency depends on understanding the purpose of a system and a general purpose system makes it harder to know what questions to ask. One important consideration is the distinction between those building the AI and those deploying the AI, as the deployment is more likely to be oriented towards a particular purpose. As these AIs are especially effective at writing content that can appear human-written, transparency is especially essential to ensure any audience understands the degree of involvement from both humans and AIs in the ideation, organization, and drafting of any text.

4. Motivations and Priorities of Builders Around AI Transparency

Despite recent advances in AI transparency training in computer science programmes, companies still see it as a 'nice to have' rather than a necessity unless there are clear regulations and standards in place.[16] Additionally, many builders lack awareness and empathy for the potential harm their decisions can cause. Our survey shows that AI transparency is therefore often overlooked by companies and leaders, leading to gaps in builders' skills and a low sense of personal accountability.

The time our respondents devote to AI transparency in work varies widely with most respondents devoting 20-50% of their time to the topic. The main motivations for working on transparency implementation are to debug and improve the AI models for higher accuracy of results. Other motivations include increasing control and agency for

[15] Wikipedia on "Clickbait"

[16] AI already underlies general regulation like data protection, copyright, competition or anti-discrimination laws. Whether developers are aware of these regulations is another question we did not ask. The answers suggest that awareness of existing regulations may not be widespread.

both builders and users of the systems, and decreasing business risk. However, the focus is often on the clients of the AI rather than the individuals affected by its use.

The avoidance of business risk after deployment in a real-life setting is another important motivation. New data sources, API changes, natural evolution and more can lead to model degradation with consequences for business operations, so this stage requires constant monitoring and iterative work on transparency to control the model's behavior. But monitoring over the entire life cycle of AI systems is not the norm for builders. Most transparency tools are used at the development stage, with little sharing of deployer and client experiences or the model behavior once the AI is used in real life.

In our interviews, we heard that a lack of business incentives and organizational prioritization plays a significant role. One builder reported that their management had a business-first approach and therefore would only enforce AI transparency if it became mandatory. This translates into a push for rapid revenue growth instead of taking the time for operationalizing transparency . Ethical discussions are allowed, but they can't lead to commercially disadvantageous decisions. Young companies are more open to implementing transparency because they can adapt more easily due to less or no legacy systems.

The survey revealed a **tension between authority and advisory roles**, with only 31 out of 52 respondents having veto power regarding and 21 having only a supporting and advisory role in AI decisions. One respondent stated: *"Though I have the ability to voice concerns, there are other measures and checks and balances and approvals in place to veto my recommendations or concerns."*

Compliance with legal frameworks as well as audits were the least cited motivations. This is expected to change with the EU's AI Act[17]and the Digital Services Act[18] coming into force, which will cement a number of transparency obligations for AI services and products as well as algorithmic recommendation systems used for large digital platforms.

As of the time of publication of this report, it is likely that the AI Act will be adopted by the end of 2023 at the earliest, with a further two years (at least, maximum three) before the AI Act comes into force. This means that AI developers should expect to comply with the Act by 2025 or 2026.

Among other obligations, the AI Act would require providers of high-risk AI systems to adopt an "appropriate level of transparency" that is not further defined[19], including

[17] European Commission, "Laying Down Harmonised Rules on Artificial Intelligence" (Date of proposal 21.4.2021)

[18] European Commission, "The Digital Services Act: ensuring a safe and accountable online environment"

[19] See Article 13 of the AI Act Draft

providing information to users. High-risk AI systems are categorized by a predefined list (Annex III) in the AI Act and include systems that pose significant risks to fundamental rights or security, such as biometric identification, credit scoring, recruitment or self-driving cars. Fines for failure to comply with these requirements can be up to 20 million euros or 4% of the total worldwide annual turnover in the preceding financial year, whichever is higher.

As well, the **AI Auditing ecosystem**[20] is gaining traction and will become increasingly relevant and might be able to fill the lack of guidance articulated by our participants. Mozilla is supporting the AI Auditing ecosystem with various projects that are covered on our blog[21].

Binding international standards, such as IEEE Standards[22] are in an early phase of exploration and implementation. European standards are also expected[23] to consolidate and might therefore become more relevant in the field.

The survey revealed the following ranking of 12 motivations **(from the most popular above to the least)** for working on AI transparency

[20] See the ecosystem fieldscan from "Who Audits the Auditors?" by Costanza-Chock, Raji and Buolamwini

[21] See Claire Pershan, "Cutting Through the Jargon - Independent Audits in the Digital Services Act" (2023) for a good introduction of auditing around DSA, or Mozilla's project on open source AI auditing tooling (OAT).

[22] The IEEE Standards Association just introduced a new Program for Free Access to AI Ethics and Governance Standards (2023)

[23] Peter Cihon, "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development" (2019), see also European Standard setting bodies CEN and CENELEC being commissioned with the creation of a draft for harmonized european standards by the end of 2024.

1. Accuracy and target goal achievement
2. Gain new insights by investigating learned prediction strategies
3. Impact assessment to avoid unwanted outcomes
4. Avoid bias
5. Justify decision to subjects and other stakeholders
6. Enable user control
7. Increase security
8. Verify generalizability of the model
9. Disclose knowable information
10. Improve system robustness
11. Compliance with ethical guidelines / internal code of conduct
12. Legal compliance / audits

Figure: Ranking of participants' motivations for transparency work. 48 respondents.

5. Transparency Tools and Methods

Explainability of AI (XAI) may be a hot topic in research, but the practical implementation of these tools and methods is often lacking. As a result, there's a significant gap in the development and use of XAI tools, leading to a general sense of confusion and uncertainty. Many developers are left without clear guidance on which tools to use and how, and lack critical implementation skills. The current state of XAI tools is worrying, with a lack of variety and poor explanation leading to a concerning level of confidence in existing solutions. To help navigate this landscape, we've compiled a list of the tools most commonly used by our participants.

Black box machine learning or deep learning models are broadly used for prediction and decision-making in sensitive areas such as credit systems, health care, predictive policing, and criminal justice, all of which have significant potential for harm. This has led to several intended and unintended issues including discrimination[24] and false accusations of (in this case welfare) fraud[25]. The evolution of methods to explain these black box models will hopefully mitigate some of these risks.

There is a need for predictive ex-ante knowledge to understand how user input will impact future outputs of the AI system under general circumstances. This can help answer end user questions like: "*What will happen if I enter a specific answer?*", "*How will my decision impact the system?*", and "*How can I get approved for credit?*"

According to one of our survey participants, another complication results from the fact "*that builders rebuild or retrain ML models in dynamic ways that are not transparent to users. This can happen every day, every hour, or even more frequently in time-sensitive industries like financial services.*" Transparency around the model's behavior, the logic involved and "AI reasoning" is therefore an important part of AI transparency in general.

We wanted to learn about the experiences of builders with a range of different explainability methods and tools to see if we can recommend a list of XAI tools and methods that work best in the machine learning industry and in real-life settings. Survey participants identified the tools below as their most used.

It's worth noting that a distinction can be made between explainability and interpretability. In Cynthia Rudin's paper "*Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*",[26] she describes the distinction, as explained in one summary[27] of the paper:

> *In **explainable** ML we make predictions using a complicated black box model (e.g., a DNN), and use a second (post hoc) model created to explain what the first model is doing. A classic example here is LIME, which explores a local area of a complex model to uncover decision boundaries.*
>
> *An **interpretable** model is a model used for predictions, that can itself be directly inspected and interpreted by human experts.*

[24] Carsten Orwat, "Risks of Discrimination through the Use of Algorithms" (2020)
[25] Amnesty International, "Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal" (October 25, 2021)
[26] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019)
[27] Adrian Colyer, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (December 10, 2019) on the paper of the same name by Rudin.

So, XAI tools are needed to explain black box models, but other models can be inherently interpretable, and thus require no special tool for explanation. It's worth noting that the interpretability of a model depends both on the model structure (logistic regression models can be easily interpretable, while neural networks generally are not) and on the interpretability of data features, the variables that exist in the data used by the system: if the features fed into the model have natural interpretations (e.g. number of loans applied for in the past year, age group, or speed of internet connection), the model may be interpretable; however, it is possible to engineer features with no natural interpretation; a model using such features will not be interpretable, regardless of the model architecture.

It is also important to ensure that explainability or interpretability supports rather than undermines accountability and human empowerment. One analysis[28] found that "explanations could be seen as blameworthy agents, obscuring the responsibility of developers in the decision-making process" and "that XAI could result in incorrect attributions of responsibility to vulnerable stakeholders, such as those who are subjected to algorithmic decisions (i.e., patients), due to a misguided perception that they have control over explainable algorithms." It is critical that explainability or interpretability be employed in the service of meaningful transparency, rather than as ends for their own sake.

It's worth noting that many simpler statistical methods, such as logistic regression, are also marketed as AI and that the upcoming EU AI Act[29] takes a broad definition of AI that includes such models. These models, when using interpretable features, are generally not "black box" and interpretability by the builders would not require the ex-post explanation tools mentioned above (with the exception of purely descriptive methods such as model cards). The comprehensibility of these models would therefore be more a procedural issue (documentation and organizing of the interpretation) than a technical one.

a. Most used explainability tools and methods

We wanted to learn about the experiences of builders with a range of different explainability methods[30] and tools to see if we can recommend a list of XAI tools and methods that work best in the machine learning industry and in real-life settings. Survey participants identified the tools below as their most used.

[28] Lima et al., "The Conflict Between Explainable and Accountable Decision-Making Algorithms" (2022)
[29] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN
[30] See Timo Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods" (2022)

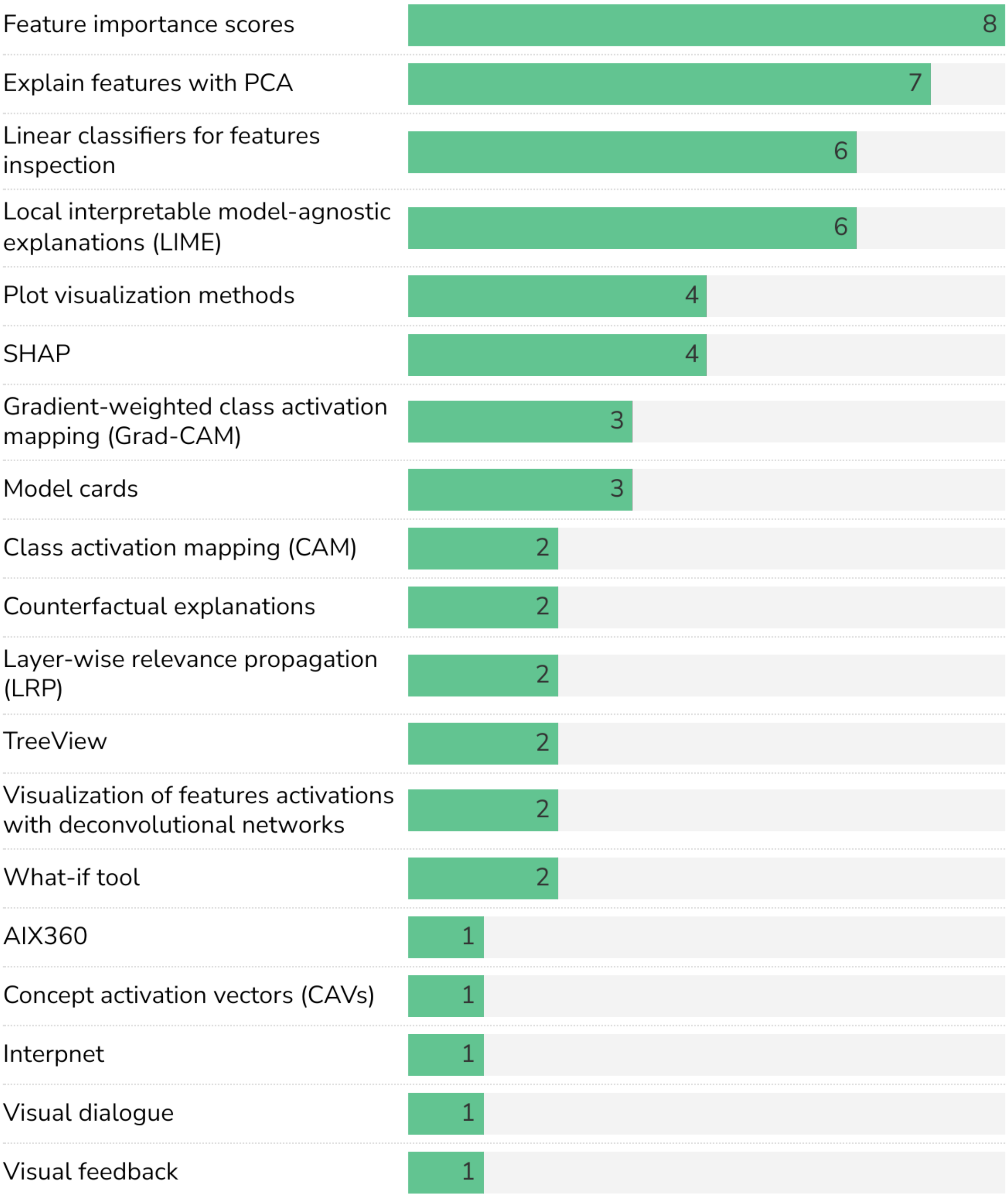


Figure: XAI tools and features in use. 17 respondents, multiple choice. Find more information about XAI tools and features in e.g. this taxonomy.

Monitoring and observability tools that relate to application performance metrics and infrastructure status are also important. Such tools can monitor error rates that may be indicative of existing biases, new biases emerging in the data, or other risks of harm. An interviewee shared that there seems to be little knowledge on builder's side regarding "*what's happening on the algorithmic and data sides of things. We don't know if our models that we put in production suffer from drifts or biases. Are there specific segments in the data for which the most underperform?*" AI monitoring tools are generally fairly robust, but expertise in their use is not always available and appropriate application of these tools is not consistently practiced.

b. Experience and critical reflection on the methods used

In the survey, we asked participants about their experiences with the tools above. Their responses suggest that the degree of usefulness of the tools varies, with some tools producing misleading or wrong explanations at times. Another critique shared is that there is a loss of information due to dimensionality reduction: if the explanation is simpler than the model (which is usually the case), it's inherently lossy. Besides this, overfitting (when noise in the training data is picked up and learned as concepts by the model), lack of user-friendliness and usability, and unclear or non-understandable explanations were mentioned several times in open answer fields.

The gap between XAI research and implementation and deployment of those tools and features in real settings was also reflected in the interviews. One participant questioned the need for such tools altogether, emphasizing that "*these post hoc techniques are typically used when the conceptual path of modeling is not understood.*"

Two interviewees said that there is a deep uncertainty about how to put AI transparency into practice and operationalize explainability. One of them stated: "*There are lots of discussions, talks, conferences, and podcasts, but there's uncertainty about the translation into business process and technical execution on the team and corporate levels. It hits how you work in teams, how you execute projects, how you onboard and roll up projects and deliver them back into the business.*"

Some attributed this lag to a certain disconnection between research and business units within an organization. Researchers often lack access to monitoring data from production environments, so they can't observe shortcomings in the model. According to one interviewee, best practices for A/B testing, version benchmarking, and machine learning ops are not consistently applied.

We also wanted to learn more about potential obstacles and hesitancy in the operationalization of AI transparency. According to our interview partners, the sheer number of tools, uncertainty about which tools to use, and the lack of diverse tools to suit different needs are common barriers to use. One example provided was the lack of fit for unsupervised machine learning beyond visualizations. We also heard skepticism about tools because they are often provided by big tech from one participant. *”I don’t feel comfortable accepting lessons or tools from large technology companies.”*

Overall, XAI methodologies, tools, and concepts are in the middle ground of technology readiness alignment; in many cases — based on the research participants’ comments in our interview and the survey — the methods seem too underdeveloped for widespread use, which seems to lead to a lack of confidence in XAI among the builders we interviewed. At this stage of technological readiness, they are still quite developmental and technical and require simultaneous advanced understanding of both the intended domain of use (e.g. medicine) *and* the specific explainability tool, and not just machine learning. Using them requires specialized training, detailed discussions, and interdisciplinary knowledge.

As an alternative to XAI, we saw a small trend in the interviews of builders recommending the use of linear models that do not require post hoc explanation methods for interpretability (beyond the statistical knowledge that will be part of domain expert users), as they remain at the level of semantic matching with linear and symbolic models. These models are often sufficient to achieve the desired results with high traceability. The use of inherently interpretable models for high-stakes decisions is a known recommendation.[31] Thus, this is not about interpretability from the laypersons’ point of view. Such AI transparency is addressed in Section 8.

c. Builders’ misgivings about increased algorithmic transparency

We asked builders if increased algorithmic transparency could lead to negative impacts. We received several survey responses that focused on the tension between increased transparency and a user’s ability to game a system in adversarial contexts:

> *“Depending on how the information is exposed, it could make it easier to develop malware that evades detection.“*
>
> *“In our fraud detection, it is problematic to provide explanations to fraud prevention actions since this is also information about how we detect fraud — which we, of course, do not want to reveal to fraudsters.”*

Another argument against increased AI transparency that lingers from early discussions [32] about the topic: depending on the organization and intended business goals, increased

[31] Cynthia Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead” (2018)

[32] Fenwick et al. “A Looming AI War: Transparency v. IP Rights | Fenwick & West LLP - JDSupra” (August 25, 2019)

algorithmic transparency could pose a threat to proprietary methodologies and technologies. While algorithmic transparency is not about the disclosure of algorithms and data but about explanations and *descriptions* of them, survey respondents expressed concerns about intellectual property too.

d. Trust in the explanations provided by explainability methods and tools

None of our 19 survey respondents for this question fully trust the outcome of model behavior interpretation methods; 63% indicated that they outright distrust the explanations, and 37% are generally skeptical or unsure.

The lack of confidence in explanatory methods revealed by our research is remarkable. It is further evidence of the gap between research on algorithmic transparency and its actual use, which is exacerbated by poor quality explanations and a lack of external pressure to accelerate adoption. Based on these clear findings, it would be premature to recommend specific explainability tools and methods.

An interesting finding was that our interviewees felt that explanations were often incorrect or irrelevant,[33] or that the people using them could not understand them properly, leading to misinterpretation. As one respondent put it, the fact that AI explainability tools are a tech trend doesn't guarantee quality. "*As with any tech trend, the problem is that everyone (only) wants to show that they are using explainability without really understanding it*."

Another interesting issue raised by the builders is that domain experts often don't admit when they have problems understanding explanations, so they don't ask developers for clarity or provide feedback that could improve the quality of explanations.

The lack of trust in XAI methods that our respondents expressed is reflected by broader industry and academic trends. As in a recent article[34] on the EU's AI Act:

> Zachary Lipton, a machine learning researcher at Carnegie Mellon University, says that "[e]veryone who is serious in the field knows that most of today's explainable A.I. is nonsense"; a 2020 study has shown that the well-known tools LIME and SHAP (which are already limited to explaining individual decisions, rather than a systems' general decision-making) can be abused to make untrustworthy models seem reliable; and a summary of the field has concluded that "it remains unclear when—or even if—we will be able to deploy truly interpretable deep learning

[33] See also Ghassemi et al., "The false hope of current approaches to explainable artificial intelligence in healthcare" (2021)

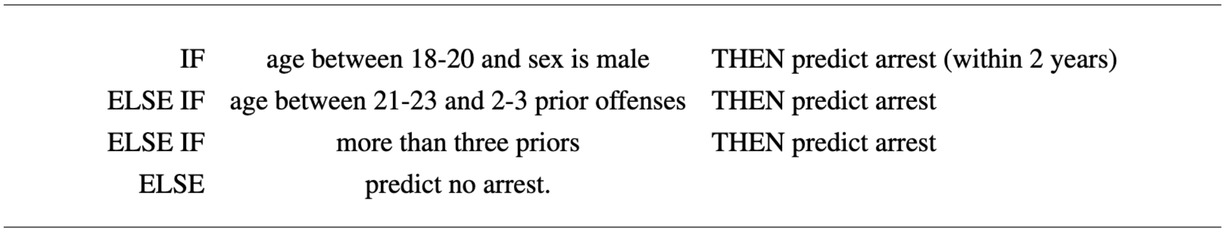
[34] Hadrien Pouget, "The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist" (January 12, 2023)

> systems [for example, neural networks].” If these measures worked, they would be useful, but for the moment they remain wishful thinking.

So there are limits to XAI tools, but there are alternatives.

As Rudin mentioned at the beginning of her paper: “*Interpretability is a domain-specific notion, so there cannot be an all-purpose definition. Usually, however, an interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, or physical constraints that come from domain knowledge.*”

The paper also provides an example of a interpretable model:



IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

This extremely simple model is actually, as described in the paper, just about as effective as the complex, proprietary, black box (and described as racially biased)[35] model COMPAS used by the US Justice System. There are of course some other cases in which more complex models are significantly more performant, but in this case, gains are minimal and come at the cost of opaque and potentially inappropriate criteria.

There is a common assumption that interpretable models are less accurate than more sophisticated black box models. This is sometimes true, but not in as many cases, or to such a degree, as many might imagine. In any case, interpretability and accuracy are both important design criteria for an AI system and while sometimes tradeoffs must be made, in many cases, an interpretable model can be found that is at least approximately as accurate as any other non-interpretable model.

As we have seen, post hoc explainability can have limited value, but interpretability by design is a useful orientation to enable meaningful transparency – especially for high-stakes deployments in which interpretability is a design requirement.

[35] Larson et al., “How We Analyzed the COMPAS Recidivism Algorithm — ProPublica” (2016)

6. Awareness of Social and Ecological Impact

Understanding the broader impacts of AI beyond algorithmic transparency (explainability) in the narrow sense[36], is critical for meaningful transparency. Careful consideration of what a system is designed to do and the extent to which it is fit for that purpose is part of the concept of meaningful AI Transparency, but so is consideration of risks of unintended consequences or the understanding of broader social impacts. Despite some efforts to reduce bias in the data and to share information about these measures, the builders of AI systems don't seem to consider communicating important factors such as the purpose of their use, the metrics used to optimize performance, and the process around assessing risk of unintended consequences for individuals and society.

a. How organizations address ethical issues that arise using AI

Our participants shared that ethical discussions are widespread during development, but they rarely extend to the creation of explicit processes to communicate common pitfalls, like discrimination or limitations of the AI system. As one survey respondent mentions:

> *"There have not been a lot of such situations. There is a focus to avoid such a scenario by checking all angles before taking the AI system to prod[uction]. But rarely do users understand the system enough and have the courage to question it to such an extent as to bring ethical issues to light."*

Risk mitigation plans appear to be rare — most interviewees and survey respondents told us that their organizations don't have a strategic plan that considers different scenarios and corresponding actions based on the nature of an incident or the domain in which they operate. They also noted a lack of appropriate feedback loops between builders and users who question their AI systems. Nevertheless, 11 out of 17 respondents to that question believe in their organization's ability to identify and address issues related to data and algorithmic harm.

b. Evaluation of impact on individuals, society, and environment

AI systems can require significant computational resources to train and deploy. This can mean massive power consumption which generally corresponds to significant carbon emissions. Many cloud providers claim carbon neutrality, but this is generally based on offsets that have been shown to have very limited legitimacy[37]. Transparency can

[36] See the wide context of AI Transparency we introduce in Section 1.

[37] Patrick Greenfield, "Revealed: more than 90% of rainforest carbon offsets by biggest certifier are worthless, analysis shows" (January 18, 2023)

include information about computational resources used, power consumed, and carbon emitted. Legislation to require this is already coming, for example, in France[38].

It is estimated that "the global tech sector accounts for 1.8% to 3.9% of global greenhouse-gas emissions. Although only a fraction of those emissions are caused by AI and machine learning, AI's carbon footprint is still very high for a single field within tech [39]." As AI is a rapidly-developing field, it is critical that adequate transparency of carbon emissions is established as a norm to allow tracking of the environmental consequences – what gets measured, gets managed.

Other algorithmic harms can also be invisible. An ex-ante assessment of possible intended and unintended harms can be crucial. This was the motivation behind this part of the survey.
Around 7 out of the 15 respondents to the first question here stated that they have conducted internal impact discussions/assessment to understand and evaluate their AI's impact on individuals and society. We don't have more detailed information about the type or granularity of the internal "assessments".

Nine out of 15 participants said that their organizations do not address the ecological impact of their AI in use. Five of them do assess environmental aspects, particularly energy consumption and CO2 footprint.

Working conditions of tech workers have been featured in the media[40] over the past years, including in the documentary The Cleaners,[41] which explored the mental health of content moderators. In our survey, 6 out of 16 respondents to the question around tech work said they are knowledgeable about the working conditions of tech workers involved in their work (e.g. content moderators and data labelers for supervised learning). Another 6 stated that they are somewhat informed, and 4 survey participants said that they don't have any information about labor conditions.

Eight out of 18 respondents to the last question of that part generally sympathize with initiatives and grassroot campaigns like #Data4BlackLives,[42] #NoTechForICE,[43]

[38] Mathieu Pollet, "New law forces French operators to disclose carbon footprint to public" (November 9, 2021)
[39] MIT Technology Review, We're getting a better idea of AI's true carbon footprint (November 22, 2022)
[40] Tech Equity Collaborative "Separate and Unequal: How Tech's Reliance on Disproportionately Diverse, Segregated, and Underpaid Contract Workers Exacerbates Inequality" (October 14, 2021)
[41] Documentary "The Cleaners" (2018)
[42] Data for Black Lives
[43] #NoTechForICE

#RecruitMeNot,[44] and #TechWontBuildIt[45]. Just as many are indifferent. Two were not aware of those campaigns, and one participant stated that "*tech should not be political.*"

c. Privacy impact assessments of data and AI

When asked if their organizations assess the privacy implications of AI models in use, 14 out of 16 survey respondents answered "yes." This is interesting but not surprising, as well-defined and prescriptive data protection rules and guidelines are implemented in several states. Privacy assessments are much more advanced and established than assessments of other societal and environmental impacts.

d. Extent to which stakeholder participation is organized

A central requirement of formal technical impact assessments is the organization of stakeholder participation during the development and deployment cycle. Nearly all (15) of the 16 survey respondents here said that they involve end users in the design, testing, and development phases of system creation. But only half of participants indicate that they conduct impact assessments (which should include stakeholder involvement).

e. Conclusion

We have seen the risks of AI systems can occur at many levels and a responsible system must consider the full range of risks. Impact assessments on AI harms and privacy impacts are important tools and engagement with stakeholders is critical. Transparency must serve individual stakeholders like subjects of the system as well as the interests of society. Early involvement of all possible stakeholders should be organized from the early stages of development and throughout the lifecycle of the deployment of the AI system, in order to better understand potential unintended harms and to avoid the high costs of ex-post compliance. Although not mentioned by survey respondents, we also note that the Thoughtworks Responsible Tech Playbook[46] features a number of proven tools for consideration of possible unintended consequences.

The importance of enabling research by academics and civil society should be considered. These communities have been at the forefront of understanding the social impact of AI systems but are often stymied by limited access to needed data or other information from the AI systems they study. As such, an essential aspect of transparency is to provide the access needed by third-party researchers to study the

[44] #RecruitMeNot

[45] #TechWontBuildIt

[46] Responsible tech playbook | Thoughtworks (tools can be found starting page 13)(2021)

social impacts of AI systems. These needs will vary depending on the application and should be understood in collaboration with the researchers themselves.

7. Transparency Delivery - Practices & Recommendations

The focus of our survey participants is clearly on the delivery of information to their key audiences, being the clients, internal developers, and decision makers (e.g. product owner, CEO). Providing appropriate information to various stakeholders seems to pose great difficulties for developers and reveals the limitations of meaningful transparency. The lack of literacy required to understand AI systems only exacerbates this challenge, making it imperative for developers to find new and innovative ways to effectively communicate how their AI works. The mismatch between the type and amount of information currently provided to stakeholders and what they believe is necessary and useful is a significant issue that needs to be addressed.

As transparency is only a means to an end, it is important to determine whether the transparency provided actually enables each stakeholder to make informed choices such as challenging and objecting to algorithmic decisions, opting out, identifying the responsible entities behind the AI, or assessing fairness.

The leading questions of meaningful transparency are:

1. Given a stakeholder's perspective and literacy, is the provided information clear and understandable?
2. Will the information satisfy the reason for their need for information, such as to know who is responsible, or to be able to challenge the automated decision, or to verify whether the decision was fair?

To us, meaningful transparency means finding the right form and depth of transparency delivery to meet a stakeholders needs, in a form that makes sense and is useful to them. When we asked, "What is the right way and degree to deliver AI transparency to different stakeholders?" half of the participants didn't answer the question.

In interviews, this question was met with the following responses: *"Million dollar question!" "The hardest question!"* and *"Very good question!"* But no one provided a concrete answer. The delivery of explanations seems to be especially difficult.

Nevertheless, the builders shared relevant insights, confirmed some trends, and revealed gaps. Discovering the right amount of transparency seems to be just as important as formulating the limits of transparency.

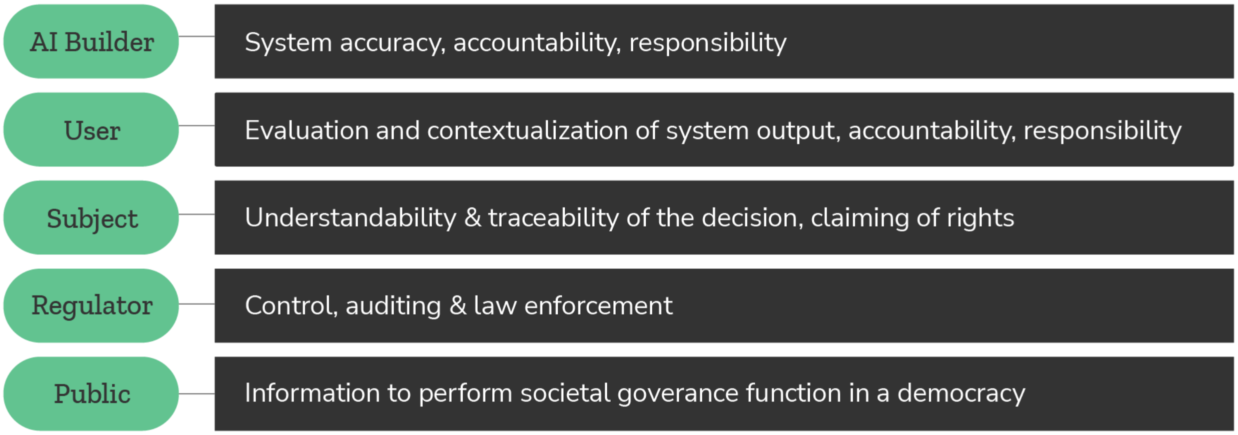
Transparency concepts are partly based on outdated solutions that have led to unsatisfactory results[47] such as too long or incomplete privacy policies. **A particular danger of too much focus on AI transparency as a goal is that accountability can be shifted to those affected by the algorithmic decision – transparency should not put an excessive burden on stakeholders.** Transparency then acts as a substitute for accountability instead of ensuring it. This fallacy, which is also known from the privacy field[48] regarding deceptive design to obtain consent, should always be taken into account.

In some cases, it may be suitable to consider delegation.[49] Some information or decisions may be too complex for some stakeholders, in which case delegates with aligned interests can act as proxies for the stakeholders. For example, car buyers are not asked to evaluate the safety of each car model themselves, but instead they delegate assessing the transparency information about car safety features to standards and regulatory bodies. Similarly, some AI transparency information may be suitable for regulators or other third parties, rather than users or subjects.

We also want to point to a specific fallacy of transparency delivery: research on people's perceptions of informational fairness — "whether people think they are given adequate information on and explanation of the decision-making process and its outcomes"[50] — and its relation to trustworthiness shows that recipients' AI literacy plays a relevant role *and* that people can be nudged into trusting AI by simply providing them a lot of information about it. On the other hand, other research[51] shows that effective transparency from an AI system can make users "more inclined to second-guess its recommendations". This is something to keep in mind when reflecting about the right amount of transparency.

Different stakeholders have different transparency needs related to their divergent interests and motivations, as detailed in the exemplary listing below.

[47] Divine Q. Agozie a, Tugberk Kaya, "Discerning the effect of privacy information transparency on privacy fatigue in e-government" (2021)
[48] Jonathan Zong, "Changing attitudes toward (mis)use of personal data" (2020)
[49] Richard Reisman, Delegation, Or, The Twenty Nine Words That The Internet Forgot (February 28, 2022)
[50] Schoeffer et al., "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making" (2022)
[51] Sara Brown, "Why employees are more likely to second-guess interpretable algorithms" (2023)



AI Builder	System accuracy, accountability, responsibility
User	Evaluation and contextualization of system output, accountability, responsibility
Subject	Understandability & traceability of the decision, claiming of rights
Regulator	Control, auditing & law enforcement
Public	Information to perform societal goverance function in a democracy

Figure: Transparency interests of selected stakeholders

In the survey, we asked builders about the transparency they *currently* deliver and how they think transparency *should be* delivered. The following insights are therefore subdivided into actual transparency delivery and recommended delivery.

Note: The following transparency provisions are purely exemplary mentions and thoughts that do not allow any conclusions to be drawn about widespread implementation. However, we feel that they provide valuable insight into nascent transparency implementations. It's also important to note a point on survey design: the questions around actual delivery were higher up in the survey, where we still had 26 to 28 participants, while the question about recommended delivery was near the end of the survey, when we had 4 to 19 participants. This explains the deviation in the number of respondents.

a. Transparency for clients and domain experts

Alongside the clients who use the algorithmic systems and who are often domain experts, other internal staff, including other developers and decision makers, also have transparency needs. So we started by asking respondents about transparency provided within their organizations. We learned that it can take the form of model cards, logs, and short textual descriptions of how the system works, created for other internal machine learning experts. In addition, implementation details and results are sometimes shared with engineering management.

Most of the transparency efforts in the organizations represented appear to focus on the design (also called development) phase, rather than the full lifecycle of the AI system, which includes the deployment and run stages. Respondents indicated that their organizations take a contextual approach to providing transparency, focusing on the development phase rather than the specific use and relating the information to the context of use. Essentially, more effort is being put into transparency efforts that benefit experts and people within the organization rather than end users.

When it comes to providing information to clients and domain experts, these are our findings.

➢ Delivery in practice

According to 26 survey respondents, transparency delivery to clients and domain experts primarily consists of real time or timely explanations and reports, sometimes supported by visualizations of the following information:

- Report on concrete examples of (previous) system decisions
- Decision decomposed by factor, plus a description of each factor
- Information about similarities of examples
- Description of dataset and model inputs and outputs
- Details on the system architecture and implementation
- Full access to raw data in their account, including AI training and model data

Some survey respondents shared that *“the explanations are typically shared via simple documentation”* and are *“sometimes presented in meetings or workshops.”* Some builders rely mainly on transparent (interpretable) models such as decision trees and provide information about the importance of features used in a model or decision. Few provide detailed explanations of different levels of abstraction or engage with stakeholders to ensure a thorough product.

Two builders told us that they only share what they consider important regarding the system to the client and end users **to detect and debug anomalies**. This is often provided to system operators in the form of model cards; diagrams and visual representations of data, system architecture. Model results are also common methods of communicating transparency. In addition, sample model decisions are sometimes shared.

Combining interpretable models with documentation is a well known approach[52] to transparency, and some of our participants also favor this approach. A few respondents are already working with explanations for stakeholders by building a chat bot to **facilitate explanation dialogue instead of one-sided information delivery.**

[52] Cynthia Rudin and Joanna Radin, Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From an Explainable AI Competition (2019)

Once deployed, a significant number of AI systems are either multiuse or can be used in different contexts by the client. So not only is the AI system itself dynamic regarding retraining and changes that might occur to the system, but the deployment context is also dynamic. This is why we asked if and how survey participants' organizations share **information on conditions for repurposing or deploying AI systems in diverse contexts** by clients or end users. Of the 14 builders who answered the question, 10 said they either don't share that information or that they are planning to do so, but haven't done it yet.

Two respondents stated:
"There is no direct access to our ML models by the client for security reasons, so repurposing is not possible.“
"We do this very strictly with caveats and assumptions and operating conditions and we would recommend this not be done at all without developer advice."

Recommended delivery

We asked survey participants to rank the following possible types of transparency delivery for clients and domain expert users. This is the ranked list of recommended deliveries from 16 participants:

1. Explanation surface/dashboard to understand why or why not algorithmic system recommends x
2. Data reporting (e.g. data sheets for datasets)
3. Explanation surface/dashboard to know when and when not to trust an outcome
4. Explanation surface/dashboard to when it succeeds or fails
5. Explanation surface/dashboard on how to opt out, override, or correct the AI system's output
6. Insights into model behavior
7. Transparency around limitations: adding occasional obvious errors to system outputs to make users aware of system limitations & model reporting (i.e. model cards)

Interestingly, the recommendations builders selected in the survey differ from their actual implementation. As an example, the highly ranked dashboards are not offered by any of the respondents' organizations. Their own ambitions of implementing stakeholder-adapted[53] approaches like *"the client needs to understand as a layman or technical person how the model works, what it does and why"* as well as *"the delivery should follow a simple explainable and non-technical approach"* and the reality of transparency delivery diverge significantly.

Builders also told us that they often use model cards, but they were ranked the lowest. The fact that they are still used may be due to their familiarity and the fact that they do not go beyond a simple description, making them the minimal option open to all who care about transparency. It's worth noting that users' AI literacy may mediate the degree to which explanations increase trust.[54] When it comes to model cards and other methods of providing meta-level explanations, there is new skepticism rising in research about "metadata maximalism."[55]

b. Transparency for individuals affected by an algorithmic decision

"Being well-informed is crucial for exercising one's autonomy and represents a vital element of transparency."[56] Decision subjects, those affected by the use of AI, need transparency that enables them to understand and contest algorithmic decisions and voice their disagreement.

In the survey, we asked builders about the transparency they provide to impacted individuals and their recommendations for ideal transparency for these stakeholders.

➤ Delivery in practice

We only received one answer about the transparency builders currently deliver here: "The end user (the person being subjected to the outputs of the system), in general cannot debug a system due to the nature of the outputs."

[53] Trustworthy AI | IBM Research

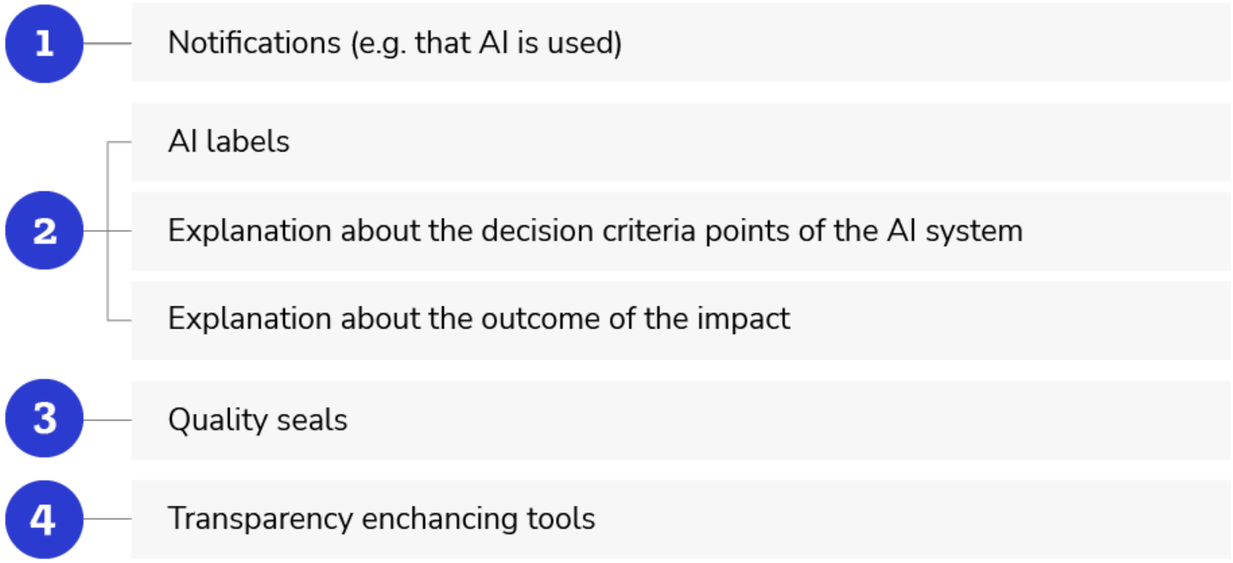
[54] Donghee Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI" (2021)

[55] Ben L. Gansky and Sean M. Mcdonald, "CounterFAccTual: How FAccT Undermines Its Organizing Principles"(2022)

[56] Ida Varošanec, "On the Path to the Future: Mapping the Notion of Transparency in the EU Regulatory Framework for AI" (2022); Buijze, A.W.G.J.; "The principle of transparency in the EU law" (2013)

➤ Recommended delivery

We asked builders to rate the following possible types of transparency delivery for this cohort of individual stakeholders (17 participants rated).



1 Notifications (e.g. that AI is used)

2 AI labels

Explanation about the decision criteria points of the AI system

Explanation about the outcome of the impact

3 Quality seals

4 Transparency enchancing tools

Other options that weren't ranked, but mentioned by participants in the open answer field:

"Always depends on the domain and context of use." "This entirely depends on the context. Which AI for which purpose, for which subjects?"
"Audits.""Clear communication channel to appeal AI decisions." "Don't use AI in isolation on subjects. Maybe don't use it at all."

One participant said:"There is no meaningful way to provide AI transparency to the subjects, because subjects will likely not have the resources to process the information. They may for individual systems, but not when every service uses an AI and provides this information."

The developers noted the potential to overwhelm subjects with the wrong kind of — or just too much — information. One survey participant said that "*more of this type of transparency could actually lead to less trust.*" The connection between this type of fatigue and loss of trust can be observed in the privacy field and the development of cookie banners.[57]

Participants didn't share a desire to address the impact of AI on society. A look at past

[57] Habib et al., "Okay, whatever: An Evaluation of Cookie Consent Interfaces" (2022)

influential AI research confirms that builder's focus is primarily on technical performance [58] and accuracy, not on user rights or ethical principles of AI.

c. Transparency for regulators and watchdogs

We asked survey respondents about the provision of information about metrics, justifications, and limitations to inform regulators and watchdogs, including journalists and non-governmental organizations (NGOs).

➢ Delivery in practice

None of the 26 participants in this part of the survey provided current methods of delivering transparency to NGOs and other interested institutions. This is consistent with our observations; for example, we know from Mozilla's YouTube Regrets[59] project just how difficult it is to uncover information about the video site's recommender system.

The lack of access to this information can be attributed to a lack of regulatory obligations. In our experience, organizations respond to legal transparency requirements regarding data privacy (e.g., in the form of internal documentation and privacy policies) and deletion requests[60] only after a corresponding law (in that case the Network Enforcement Act in Germany) has come into force.

One builder stressed that regulators need more transparency than watchdogs and pointed to the "*sensationalist nature*" of how AI issues and transparency topics are handled in the press.

➢ Recommended delivery

We asked builders who took the survey to rate the following possible modes of transparency delivery for regulators and watchdogs; 17 participants provided the following ranking.

[58] Birhane et al., "The Values Encoded in Machine Learning Research" (2022)
[59] Mozilla's "YouTube Regrets" Report (2021)
[60] As in this transparency report list (as of February 2023)

1. Access to all outputs (decisions, rankings etc.) made by the AI system
   Algorithmic impact assessment audits (before/after deployment)
2. Incident reports during life cycle of AI system
3. Model reporting (i.e. Model Cards)
4. API to enable audit during the whole or parts of the AI-lifecycle.
   > Input (train dataset when developing or user data when using the system) and out put data (at least) should be given to all stakeholders (representative at least)

d. Transparency for general public/society

➢ Delivery in practice

Questions about providing information to society at large were not very fruitful: none of our survey participants are actually working on providing AI transparency or other information about the algorithmic system to the general public.

➢ Recommended delivery

Fifteen builders rated the following possible modes of transparency delivery for the general public:

1. Transparency policy
2. Open registry (administered by authorities)

Two survey participants stated that there is no meaningful way to provide AI transparency to the general public, but they each named different reasons for their thinking:

"I don't see demand from the general public." *"It will be whitewashed and unintelligible by choice."* One builder proposed a *"mitigation of harm plan"* that included a "*report on*

compensation strategies."

One interviewee voiced concern about the potentially harmful consequences of increased transparency in terms of fraud and manipulation. Another mentioned fear of public scrutiny and loss of reputation. Find further statements from our participants on the subject of transparency and trust below.

"Transparency is one solution to the trust problem, but not necessarily the best solution. It depends on what type of trust you want to achieve. For public trust in AI, we should aim for institutional trust, where quality seals are one part, and good regulations and a justice system that enforce them are another. For trust in an AI provider, the institutional trust may need additional organizational trust. These are different problems with different solutions that need to take the resources to process transparent information into account, otherwise it will lead to resignation and distrust."

"Back to basics used in government and statistics — they are all there, they are all suitable, and they all work. It's common sense for the most part; ethics and transparency is built in, not an afterthought."

One participant suggested *"providing full descriptive and technical explanations via a combination of API documentation, user and developer manuals, and design schemas."* He stressed that they should be simple and easy to understand.

8. Ranking Challenges For Greater AI Transparency

We asked survey participants to rank issues they see in the implementation of AI transparency. We also asked about this topic in interviews; this is the area where we received the most overall responses. The presence of a number of problems could explain the reluctance to use transparency tools and provide insight into why AI transparency as a whole has not yet arrived as a standard requirement in organizations.

The biggest challenges and obstables to AI Transparency

1	No clear guidance how to choose among explainable AI methods
2	The outcomes (explanations) are too ambigious (unclear) and/or still too complex
3	Decrease of development speed
4	Explanations are not correct Lack of clear objectives/KPIs to build an incentive around transparency deep dive Explanation of AI is not part of the education of ML professionals
5	Absence of strandardized evaluation methods Cost of transparency implementation Lack of buy-in from CEO/Lead Lack of clear accountability for the transparency topic Transparency measure could enable malicious users to increase capabilities and performance of undesirable systems
6	Lack of resources Lack of whistleblower protection for employees
7	Does not align with our work practices Out transparency efforts are pure ethics washing
8	Lack of internal expertise on how to use explainability techniques

Figure: Ranking of the biggest challenges in moving toward greater AI transparency. 16 respondents, multiple choice.

The survey results already indicate a problem of trust in the use of explanatory tools. Our interviewees confirmed these reasons and identified others. They also expressed additional concerns and barriers, which are outlined below.

An interesting paradox is that our participants cited "lack of regulation" as a reason for the missing external incentives and guidance, despite the fact that there is already extensive regulation of AI, such as data protection, copyright and competition law, or anti-discrimination law. In the course of our research, we gained the impression that there may be a lack of motivation to implement transparent AI both within organizations and outside them, with nearly no expressed interest from end users or the public. This is likely due to the lack of enforcement of existing laws, the lack of harmonized standards and little pressure or interest from external stakeholders to justify additional developmental efforts. Some builders stated that it should not depend on an individual builder's will, but on governmental regulation. At the same time, one interviewee expressed concern regarding the ability of governments to draft successful regulation and enforcement; he suspected that a lack of expertise and dedicated roles at the government level might be blockers.

Several interviewees cited a lack of maturity and skill set gained by education on the topic of AI transparency as reasons for the lack of builder participation in this work. This self reflection echoes Stanford philosopher Rob Reich's statement that the AI profession is still like a "late-stage teenager."[61]

AI transparency was not covered in the education of many of today's machine learning experts and data scientists. These days, there are electives on the topic and lots of research, so some students are coming out of universities with at least an awareness of some aspects of AI interpretability for more trust and fairness. Entering the workforce makes the need for transparency more obvious, if only for debugging purposes[62], but we are still in an early stage of a shift to AI organizations universally considering the topic.

In general, we didn't encounter resistance to gaining more insight into model behavior, but there appears to be a lack of awareness of the full picture. It seems the blockers are minimal exposure and business prioritization. Research[63] reveals a lack of empathy and solidarity toward impacted individuals, with few AI developers questioning if what they are building is ethical. Questions like, "If the model is wrong, does it have a harmful effect?" are rarely asked. The same goes for questions about the ecological impact of AI.

[61] Edmund L. Andrews," Rob Reich: AI Developers Need a Code of Responsible Conduct" (June 22, 2022)
[62] Bhatt et al., "Explainable Machine Learning in Deployment" (2020)
[63] Ramya Srinivasan, "The role of empathy for artificial intelligence accountability" (2022)

Stay in touch & register for our new project

Exciting news! Our transparency project is moving forward and we're delving into the specifics of Article 13 of the upcoming AI Act and the requirement for AI labeling/declaration (*e.g. watermarking of AI-generated content*) under Article 52.

If you would like to participate in our research, know about existing **AI labeling projects**, or attend our **upcoming workshops on implementing AI transparency**, send an email to Ramak Molavi Vasse'i.

Don't miss this opportunity to learn and contribute to our mission. Also, join our Trustworthy AI Slack channel and learn more about the Mozfest working group.

Annex: Research Methodology

We chose a research method that reflects the mission of our meaningful AI research, which is to bridge the gap between ongoing, in-depth research on explanatory methods and AI transparency in practice. It is therefore an exploratory action research with a larger scope than usual to also capture trends and evaluate next research and action steps. Rather than focusing on a narrow scope, we used this research to unfold AI transparency to the most extent possible.

We employed mixed methods research for this project. The survey we conducted was anonymous. We collected quantitative data from multiple choice fields, rankings etc. and qualitative data from open answer fields. We gathered even more qualitative data for our subsequent analysis from the interviews we conducted. We merged data from open fields in the survey and the interview data and then applied a qualitative analysis method on it.

We reached out to participants through searches on Linkedin, Twitter, Github, external Slack channels, and other resources, and primarily targeted those who indicated they were active in AI transparency or AI assessment on the web. We also asked AI ethics circles and machine learning experts in our network if they knew of colleagues working on AI transparency. In addition, we reached out to startups that offer AI transparency solutions as a service. Look at this part above. to learn more about our participants.

We sought to include a diverse group of builders in terms of geography and gender. Survey participants had the option to indicate their location, which nine of them did. Of those who shared, we had representation in Europe (England, Scotland, Netherlands, Germany, and Cyprus), India, and the United States. We are pleased to have participants from three continents. However, this also shows the limitation of our results: this survey cannot be representative of the entire builder population or of diversity in the field.

Acknowledgement

We're excited to present this report on AI transparency, a collaborative effort led by Ramak Molavi Vasse'i at Mozilla's Insights Team as research lead, with co-authorship of the report from Jesse McCrosky (Thoughtworks) and editing by Kenrya Rankin.

Tracy Kariuki led engagement efforts together with Audrey Hingle, Kevin Zawacki, as well as Linda Horiuchi and Linda Pfaff from Thoughtworks.

We're grateful for the feedback provided by Stefan Baack, Christian Bock, Lindsay Dearlove, Maximilian Gahntz, Tomas Heiskanen, Solana Larsen, Crystal Lee, Jackie Lu, Kasia Odrozek, Neha Ravella, Becca Ricks, and Amber Sinha and D'Andre Walker. We thank Kasia Odrozek and Ashley Boyd for informing the direction of our AI transparency research.

We'd also like to thank Kristina Shu, Rebecca Lam, and Nancy Tran from Mozilla Foundation's design team for their support in creating this report, especially Nancy who went above and beyond to make it happen.
Thanks to Arafath Ibrahim for creating the stunning artwork for this report.

A special thank goes to Bogdana Rakova for her invaluable industry expertise and creative mind, which supported us from the very beginning of the research design process to the final report. We also appreciate Témì Popo's contributions, as they provided us with valuable insights from the trustworthy AI working group to enrich our research.

We want to thank Chhavi Chauhan, Sophia George, Marie Oldfield, Graham White and Jakub Zavrel who generously contributed their expertise to this project.

We are grateful to our colleagues across civil society who provided input, guidance and feedback on this report including Borhane Blili-Hamelin and Paola Lopez.

Finally, we want to express our gratitude to the 50+ builders who shared their experiences and challenges with us through our anonymous survey. Their contributions

were the foundation of this report, and we thank them for their open and valuable deep insights.