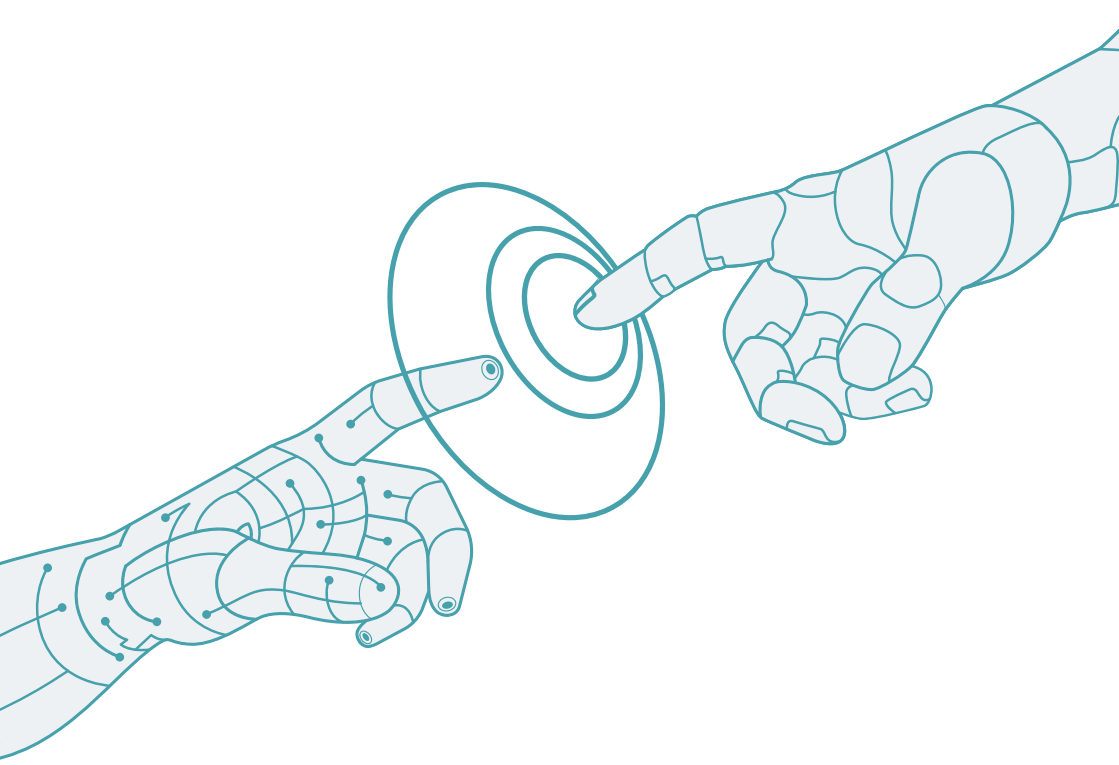
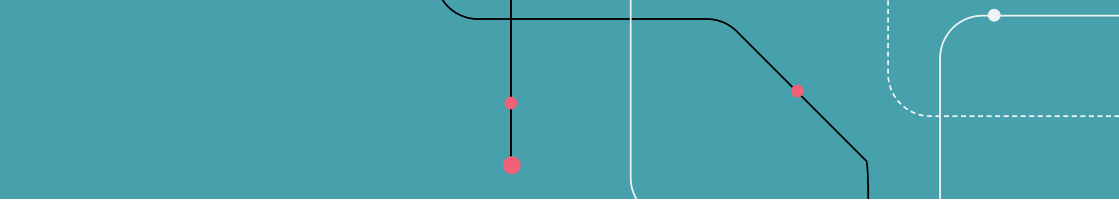


Three lenses to help design better, socially responsible AI

Jesse McCrosky,
Thoughtworks Finland

/thoughtworks



Introduction

Technology shapes our society in ways both intended and unintended. Before televisions came into our lives, our living rooms were arranged differently than they are today. Those that invented, produced and sold TVs were not setting out to rearrange our living rooms – this was an unintended consequence.

Artificial intelligence (AI) presents unique challenges to developing and deploying technology responsibly. The Thoughtworks Responsible Tech Playbook offers valuable tools to anticipate and mitigate harms from technology broadly speaking, but AI presents unique risks and its complexity demands new approaches.

AI Transparency

AI has the potential to transform society radically. That means we need to ensure discussions about this technology and the sort of society we are creating aren't restricted to small groups of people at large, powerful companies. We must all be recognized as stakeholders and given a voice. Our public institutions in particular — regulators, researchers, civil society and journalists — all need to be heard in the conversation about managing the risks of AI.

You can't have responsible AI without accountability. And you can't have accountability without transparency. Given its importance, AI transparency is today a subject of active research[1] in a range of organizations. The solutions proposed vary widely. Some work, for example, advocates sharing of code or model weights. Elsewhere, there have been attempts to develop explainability tools, intended to help stakeholders understand why an AI has made a particular decision. However, despite good intentions, such approaches are somewhat naive. They have serious limitations that need to be reckoned with. There is much more to transparency than open code or explainability.[2] Meaningful transparency means understanding the needs of diverse stakeholders and making sure that information is always actionable. According to Caroline Sinders, a machine-learning-design researcher and artist, "transparency consists of three integral pieces: legibility, auditability, and

1 For example, see our report AI Transparency in Practice, published collaboratively by Mozilla and Thoughtworks

2 Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019)

impact-ability."[3] Meaningful transparency empowers people; the wrong sort of transparency may actually decrease the sense of accountability for the developers or deployers of the AI — AI systems that provide explanations for their decisions can be seen as "blameworthy agents, obscuring the responsibility of developers in the decision-making process."[4]

A design orientation

Meaningful transparency is more than just explainability. In his 2018 paper, "The fallacy of inscrutability", Joshua Kroll rejects the argument that AI systems are too complex to understand and argues that instead of trying to generate explanations of low-level behavior, what really matters is understanding the high-level design and "operational goals and [...] their inputs, outputs and outcomes."[5] To understand a ML system, we need to understand what it's designed to do.

Tools like speculative design allow us to critically examine the potential social impact of a system. However, there are also more concrete design elements that warrant analysis. Data and metrics, for example, are both fundamental to an AI system; they operationalize important aspects of the system's design.

Metrics express what the system works to accomplish. Many unintended consequences of AI systems are byproducts of the metric for which the system is designed to optimize. Online platforms that optimize for content engagement, for instance, can amplify clickbait and inflammatory content. Similarly,

---

3 Sinders. When can we call machine learning 'transparent'? New_ Public magazine
4 Lima et al., "The Conflict Between Explainable and Accountable Decision-Making Algorithms" (2022)
5 Kroll. The fallacy of inscrutability. Phil. Trans. R. Soc. A

ecommerce systems optimizing for sales can incentivize unhealthy and unsustainable consumer behaviors.

Not dissimilar to metrics, data also expresses a particular view of the world. It doesn't exist naturally; it codifies what someone, somewhere thinks is important. This means the data a system solicits, collects and infers necessarily has a perspective. The nature of this perspective has important implications for privacy and basic human dignity and autonomy. Debates over privacy and ownership of data aside, processing data is never a neutral act.

Further, the provision or consent for processing of data is often driven by massive power and information differentials between those deploying AI systems and those using them. For example, deceptive design patterns are intended to deceive or manipulate users of a system. We will see how our framework can prevent some applications of deceptive design.

It's worth noting that we are publishing this ebook in a time when general-purpose AI systems, like ChatGPT, are seemingly everywhere. General-purpose systems present unique challenges to design analysis because, as the name suggests, they lack a singular purpose. However, many of the tools are still applicable, especially as general-purpose systems are built into products with more specific purposes.

Thoughtworks' AI Design Alignment Analysis Framework

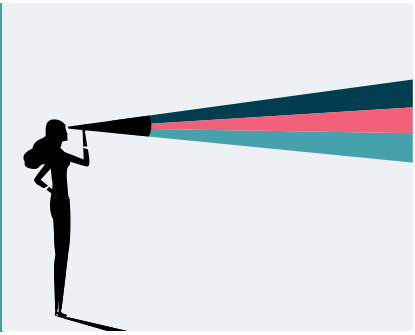
To ensure that the design of an AI system is aligned with social responsibility, we clearly need to be able to assess how an AI system has been designed. That's why we've developed what we call the AI Design Alignment Analysis Framework. This framework consists of three lenses to analyze important elements of AI systems. They are:

- **Technical function**: what does the system actually do?
- **Communicated function**: what do developers or deployers say it does?
- **Perceived function**: what do users of the system believe it does?

Each lens is useful individually, but the power of this particular framework comes from using the lenses together. By doing this, it can help us identify misalignment in a given system. Instances of misalignment are always significant; they should be viewed as a signal for failures of responsibility.

This tool can be used for multiple purposes. It can help with the audit of existing systems, or be used to guide the development or deployment of new systems in a responsible manner. To be clear, this tool alone isn't a holistic solution to AI responsibility; other tools and approaches are needed and used alongside these three lenses. Our forthcoming Responsible AI ebook will discuss responsibility more broadly; this ebook simply presents one novel approach.

The three lenses



Each lens is important. Let's dive deeper into what each one actually means and how it plays a part in the overarching framework we're proposing.

Technical function

The "technical function" lens draws our attention to the precise characteristics of the AI artifact. It encourages us to look closely at:

- Model architecture
- Optimization metrics
- Training data characteristics
- Feature engineering
- User interface and user experience design

These elements best express the design intent behind a technical system. While it's true that every line of code, every piece of data and every text, visual or audio asset can contribute to the technical function of a system, it's important to avoid getting lost in complexity or detail. Just because something is complicated doesn't mean there is no way of analyzing it or thinking critically about how and why it has been put together in the way it has.

Consider a recommender system that recommends videos. An important design choice might be optimizing for watch time — this means the algorithm "chooses" videos that are the most likely to keep users watching as long as possible. Such functionality is expressed in the optimization metrics chosen by the designers.

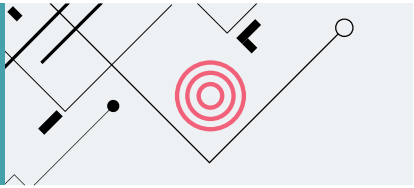
As another example, the design choice to collect and process rich data is also usually a choice to implicitly infer sensitive characteristics like gender or ethnicity[6], and thus, unless mitigated, constitutes a design choice to create a discriminatory system.

The technical function can be assessed in many ways. Ideally, details of training data, model architecture and objectives, metrics used and model source code are consulted. These can be complex and time-consuming to analyze, so when trustworthy documents are available, reading design documents, model architecture descriptions and dataset and model cards can be a good solution. And of course audits of the system itself are powerful tools.

---

6 Any adequately rich data will tend to act as a proxy for many personal characteristics. I write more about this in an article: The Inherent Discrimination of Microtargeting

Communicated function

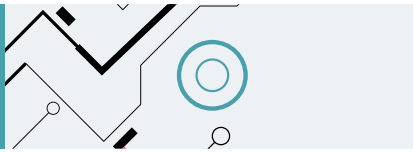


The communicated function of a system refers to everything that is communicated to everyone external to the development and deployment of the system. They could be users, regulators or even other stakeholders inside the same organization. We can understand the communicated function through a diverse range of resources, including:

- Onboarding material
- Privacy policies
- Public documentation
- In-product help or other information
- Product marketing materials

The communicated function is what the developer or deployer of the system is *telling* the world that the system is designed to do. It can be evaluated by analyzing the materials described above.

Perceived function



The perceived function is what users, subjects, or other stakeholders of the system *think* the system is designed to do.

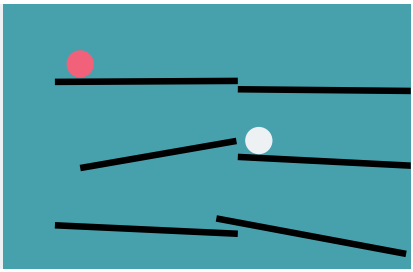
It can be evaluated through focus groups, surveys and interviews with real users and subjects. We would contend that a system in which subjects are not aware they are interacting with an AI is

always irresponsible, but as social norms evolve, the possibility of such systems may increase, in which case subjects can still be asked how they would feel about the interaction, whether or not they are aware an AI is involved. In the case of a system not yet available to the public it can be productive to do a study with potential users or subjects, reading through the information available and then discussing their perceptions of the system's function.

Interpreting misalignment

Where there's misalignment between these three lenses, there's often a failure of responsibility. As we explore below, responsible design (and basic honesty) demands that the true design of a system reflect what is communicated and understood about the system. Let's consider some examples below.

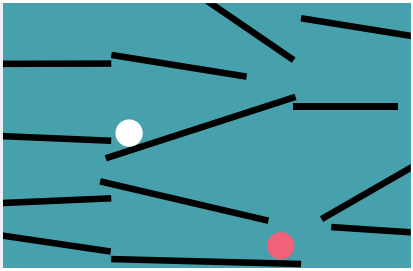
Misalignment between communicated and perceived design



This sort of misalignment is generally related to deceptive design or a lack of transparency. Take, for example, an opaque or misleading privacy policy. If the policy fails to clearly communicate how a user's data is used — or does so in a way that is evasive, complicated or ambiguous — it will lead to misalignment. The policy form itself might lead to a perception of due diligence and legality, but the content inside it does not

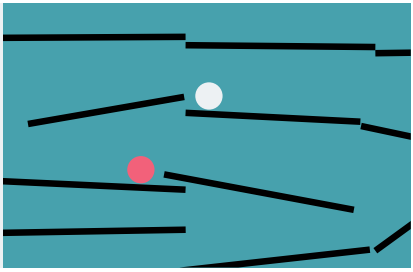
really align with the perceptions the policy creates. In this case, the responsible solution is to ensure communication is always clear. The user's understanding of the policy is the responsibility of the developers and designers.

Misalignment between actual and perceived design



This is deceptive design in the most obvious and explicit sense. Users are being misled, or, at the very least, are not adequately informed as to what the system is actually designed to do. The power and knowledge asymmetry between the users and the builders of an AI system is particularly relevant here, as special care and effort may be needed to ensure that users can meaningfully understand what may be a very complex system. In some cases, users cannot be expected to meaningfully understand a system to the degree necessary to engage with it safely; in this case, industry standards or regulations are needed to protect the best interests of vulnerable stakeholders. Even if existing regulations have not yet caught up, handling these cases appropriately is the responsible choice and mitigates future regulatory risk.

Misalignment between actual and communicated design



AI systems can be very effective at concrete goals, usually operationalized through metrics. For an AI to be able to evaluate its success and learn, the metric must be something that can be directly measured. In many cases, the purported goal of the system, for example, to serve content that the user will find interesting, does not exactly match the metric used, for example whether the user clicks on content or not. Such misalignments often create unintended consequences through a mechanism known as Goodhart's law — the idea that when a measure becomes a target it ceases to be a useful measure. We will go into much more detail on that in our upcoming Responsible AI ebook. For now, we consider in this class examples such as a video sharing site that claims to recommend videos that a user will find interesting and enjoy, when it's really designed to show videos that will simply maximize the amount of time that the user spends on the site – possibly amplifying hateful or extremist content.

Conclusion

We have seen that understanding AI is as simple as understanding what it is designed to do. The Thoughtworks AI Design Alignment Analysis Framework can help us ensure that our design is aligned with social responsibility.

About the author

Jesse McCrosky
Head of Sustainability and Social Change and Principal Data Scientist

Jesse is Thoughtworks' Head of Sustainability and Social Change for Finland and a Principal Data Scientist. He has worked with data and statistics since 2009 including with Mozilla, Google, and Statistics Canada. With Thoughtworks, Jesse is helping our clients build socially responsible AI systems, including new solutions for sustainability.

His approach to the intersection of tech and sustainability is broad, including greening-of-tech, greening-by-tech, and how technology can support the social alignment needed to tackle the climate emergency.

Jesse lives in Helsinki with his wife and two daughters.

Thoughtworks is a global technology consultancy that integrates strategy, design and engineering to drive digital innovation. We are 12,000+ people strong across 50 offices in 17 countries. Over the last 29+ years, we've delivered extraordinary impact together with our clients by helping them solve complex business problems with technology as the differentiator.

thoughtworks.com

/thoughtworks